

The Industry Supply Function and the Long-Run Competitive Equilibrium with Heterogeneous Firms*

Ignacio Esponda
(WUSTL)

Demian Pouzo
(UC Berkeley)

January 2, 2017

Abstract

The theory of long-run competitive equilibrium (LRCE), first developed by Marshall in the 1890s, has had a profound influence on our understanding of competitive markets. While Marshall referred to the notion of a representative firm, the identity of this firm is generally unclear, as the theory has focused on the case where all firms in the industry are identical. Using Hopenhayn's (1992) model of competitive industry dynamics, we extend the theory of LRCE to the case of heterogeneous firms. We show that, under certain conditions, the (long-run) industry supply function with heterogeneous firms exists and can indeed be characterized as the solution to the minimization problem of a "representative" average cost function, as originally envisioned by Marshall. As an application of the importance of accounting for heterogeneity, we show that maximal surplus is not maximized in an LRCE and that the only way to approximate the maximal surplus with a linear tax is to tax all profits and subsidize all losses.

*Esponda: Olin Business School, Washington University in St. Louis, 1 Brookings Drive, Campus Box 1133, St. Louis, MO 63130, iesponda@wustl.edu; Pouzo: Department of Economics, UC Berkeley, 530-1 Evans Hall #3880, Berkeley, CA 94720, dpouzo@econ.berkeley.edu.

Contents

1	Introduction	1
2	The model and main result	6
2.1	Setup	6
2.2	Long-run competitive equilibrium	9
2.3	Long-run industry supply function	10
2.4	Main result	11
2.5	A simple example	14
3	Proof of Theorem 1	18
3.1	Part 1. Minimization of \bar{AC}	18
3.2	Part 2. Clearing of input market	22
4	Maximal vs. equilibrium surplus	23
4.1	Planner's problem, solution, and comparison to LRCE	23
4.2	Optimal taxation	27
5	Conclusion	31
A	Appendix	32
A.1	Preliminary lemma	32
A.2	Proof of Lemma 2	35
A.3	Proof of Lemma 3	36
A.4	Proof of Lemma 4	38
A.5	Proof of Lemma 5	46
A.6	Proof of Proposition 2	47
A.7	Proof of Proposition 3	48
A.8	Proof of Proposition 4	49
A.9	Proof of Proposition 5	50

1 Introduction

The theory of long-run competitive equilibrium (LRCE) is one of the workhorse models of modern microeconomics. Currently taught in introductory courses, it is hard to overestimate its influence on people's perception of markets. The theory originates from Marshall's pathbreaking contributions in the late 1800s, including his *Principles of Economics* (1890). One distinguishing feature of Marshall's theory is his conceptualization of the (long-run) *industry supply function*. In Marshall's view, a movement along the industry supply function entails changes in other variables, most notably input prices. A question ensued as to whether such a reduced-form conceptualization is even possible.¹

Pigou (1928), Viner (1953)[1931] and others subsequently formalized Marshall's notion of LRCE. The latter author, in particular, is credited for popularizing the typical diagram taught in introductory courses and reproduced in Figure 1. The figure represents an industry where all firms are identical and characterized by the marginal (MC) and average (AC) cost functions depicted in the left panel. In an LRCE, price is at the minimum point of the AC function, p^e , and so aggregate quantity is given by the demand function evaluated at that price, Q_0^e . Suppose that there is a shift of the (inverse) demand function from P_0^d to P_1^d in Figure 1. In the short run, as the number of firms stays fixed, price and quantity increase from the original LRCE at point A to the new short run equilibrium at point B , a movement that occurs along the short-run supply function S_0 . But then firms make positive (economic) profits, and these profits attract additional firms into the market. In the long-run, the new LRCE is at point C , where all firms make zero profits at price p^e and the aggregate production increases to Q_1^e . Thus, the (long-run) industry supply function, S_{LR} , is horizontal at the minimum of the average cost function. More generally, if input prices were to, say, increase as aggregate production increases, the industry supply function would be increasing.

A second distinguishing characteristic of Marshall's analysis is the notion of a *representative firm*. Marshall recognized that there are *different* firms in an industry, and informally focused his analysis in terms of an *average* firm. The notion of a representative firm, however, turned out to be controversial, and subsequent formalizations focused on the case where all firms are identical in a long-run competitive equilibrium.

¹Barone (1992)[1894] provides an early example where Marshall's supply curve may not be well defined.

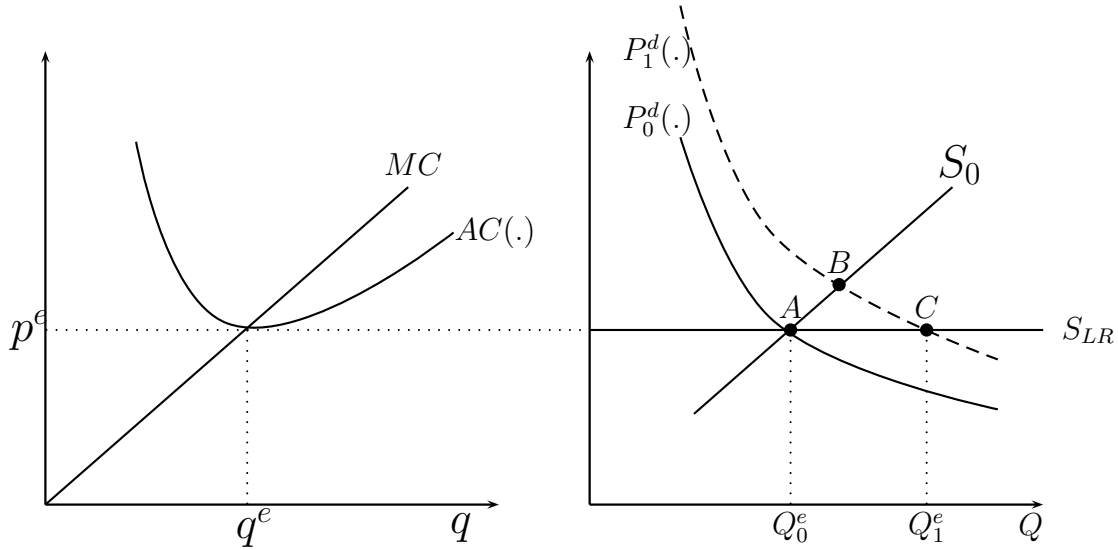


Figure 1: Textbook model of long-run competitive equilibrium.

Viner (1953)[1931], pg. 222, justifies this view:

“If there are particular units of the factors which retain permanently advantages in value productivity over other units of similar factors, these units, if hired, will have to be paid for in the long-run at differential rates proportional to their value productivity, and if employed by their owner should be charged for costing purposes with the rates which could be obtained for them in the open market and should be capitalized accordingly.”

Viner’s argument, which is the classical argument for restricting attention to identical firms, may justify why firms do not make rents in the presence of markets that bid up the price of advantageous factors, such as exceptional managerial ability. But the argument does not imply that firms with different technologies or productivities cannot coexist in equilibrium. A realistic feature of an industry is that low-productivity firms can potentially become high-productivity firms and vice versa. This feature implies that equilibrium will be characterized both by coexistence of heterogeneous firms and turnover (entry and exit), and it does not seem appropriate to exclude these realistic features from a theory of LRCE.

Our objective in this paper is to go back to Marshall’s original motivation and to extend the classical theory of LRCE to the case of heterogeneous firms. Fortunately, we don’t have to formulate a new model, since Hopenhayn (1992) actually introduced and studied a model of competitive industry dynamics where firms’ productivity evolves over time and exit and entry is an equilibrium phenomenon. We take the steady-state

equilibrium in Hopenhayn’s model as the natural extension of the theory of LRCE to the case with heterogeneous firms. [Hopenhayn \(1992\)](#), however, did not link his work to the early theory on LRCE, and our contribution is to fill-in this gap.

Our main result is that, under certain conditions imposed by [Hopenhayn \(1992\)](#) to guarantee uniqueness of equilibrium, the (long-run) industry supply function with heterogeneous firms exists and can indeed be characterized as the solution to the minimization of a representative average cost function, as Marshall originally envisioned. The standard textbook case, depicted in [Figure 1](#), is just a special case where there is no firm heterogeneity.

There are several reasons to care about this result. First, it formalizes Marshall’s original motivation of a representative firm and of the industry supply function in the presence of heterogeneous firms. Second, it provides a connection between the early literature on LRCE and the modern literature on industry dynamics (to be reviewed below). Third, it makes the model of LRCE with heterogeneous firms accessible to a larger audience (in particular, the example in [Section 2.5](#) conveys much of the intuition and can be taught in introductory courses). Finally, it helps highlight that some famous principles of competitive markets are not robust to the inclusion of firm heterogeneity.

We illustrate the last of the above points by showing that, in contrast to the case of identical firms, aggregate surplus is not maximized in an LRCE with heterogeneous firms. To illustrate, consider an industry with an infinite number of potential entrants who differ only in their unknown fixed costs. Once a firm enters, it learns its fixed cost, which becomes sunk, decides how much to produce, and finally decides whether or not to exit the industry. There is also a positive probability that firms have to exit for exogenous reasons. In an LRCE, price must be such that entrants make zero net-present profits in expectation. Thus, while there are both low and high cost firms producing in equilibrium, low-cost firms decide to stay and high-cost firms exit the industry. The selection in exit implies that, in steady-state, the proportion of low-cost firms is higher than the ex-ante proportion anticipated by firms that discount the future. Therefore, ex-post or equilibrium aggregate profits are strictly positive in the industry.

Aggregate surplus could be increased by increasing the number of equilibrium firms and lowering the price to the point where equilibrium aggregate profits are zero. At this point, however, potential entrants would not want to enter the industry

ex-ante. We show that the surplus-maximizing allocation can be achieved with a simple linear tax and subsidy to firms that remain in the market, but that it requires taxing all profits and subsidizing all losses. This is hardly a result that comes to mind when envisioning a competitive, long-run equilibrium, but it is a straightforward consequence of the facts that the planner cares only about long-run equilibrium surplus (while firms discount the future) *and* that firms are heterogeneous.

The current paper links the classic theory of LRCE, which does not explicitly model dynamics, with the modern literature on *competitive* industry dynamics. The latter literature begins with [Lucas \(1967\)](#), who studies a model where firms dynamically adjust their capital at a cost and develops a supply theory in which adjustments to demand shifts are staggered over time. [Lucas and Prescott \(1971\)](#) develop the first theory of dynamic competitive equilibrium where demand is stochastic and firms adjust their capital stock at a cost. One of the key contributions, now standard in economics, is to impose that firms have correct (i.e., “rational”) expectations about future prices. The theory, however, assumes that firms are identical and that there is no entry and exit.

Subsequent developments incorporated both firm heterogeneity and entry and exit, at the expense of no longer studying the dynamics of capital accumulation. [Jovanovic \(1982\)](#) develops the first of such models. Each period, a firm independently draws a productivity shock from a distribution that depends on an unknown productivity type. Firms have different productivity types and, as they learn their own type, more productive firms stay and less productive firms enter.

The main focus of these previous papers was to study the dynamic evolution of a competitive industry, not the steady state. Consequently, all of the interesting action happens outside the steady state. In particular, there is no entry and exit in the steady state of the models of [Lucas and Prescott \(1971\)](#) and [Jovanovic \(1982\)](#).

[Hopenhayn \(1992\)](#) is the first paper to consider a model with both heterogeneous firms and entry and exit in the steady state. In contrast to Jovanovic’s model, firms know their productivity types, but productivity types evolves randomly in such a way that firms that have a low productivity today can have a high productivity tomorrow and vice versa. As mentioned earlier, this is the model that we will use to formalize Marshall’s idea that the LRCE of a competitive industry is characterized by the cost function of a “representative” firm.²

²For a model that incorporates capital accumulation to Hopenhayn’s competitive framework,

Our approach to formalizing the industry supply curve falls in between the two classical approaches in the literature. (We base our discussion on the insightful historical account of these two classical approaches provided by [Opocher and Steedman \(2008\)](#)). In the first approach, which originates from Marshall and is formalized by [Pigou \(1928\)](#) and [Viner \(1953\)](#)[1931], the cost function of a firm depends both on individual output and aggregate output. The inclusion of aggregate output captures, in a reduced-form manner, the idea that expansion of the industry may affect input prices and affect the costs of production. Many textbooks continue to use this shortcut, most likely for pedagogic purposes, and we could have proceeded in this manner as well. This approach, however, was criticized by subsequent literature (e.g., [Kaldor \(1934\)](#), [Allen et al. \(1938\)](#), and [Hicks \(1946\)](#)) that insisted for microfoundations of the industry supply function. In particular, the second approach defines the industry supply function simply as the sum of all the individual supply functions, where the latter are solely a function of the price of the product, and all other variables (e.g., input prices) are held fixed. We follow the spirit of the second approach, in the sense that aggregate behavior arises from optimal individual behavior, and we study market clearing simultaneously in all markets, including input markets. But we strive to achieve Marshall’s objective, which is to define an “industry supply function” in the price-quantity space and where movements along that function already take into account how variables in other markets (e.g, input prices) are being adjusted in equilibrium. We show that a microfounded Marshallian approach is feasible provided that we make the assumptions that [Hopenhayn \(1992\)](#) made to ensure existence of a unique industry equilibrium.

In [Section 2](#), we present the model, the main result, and an example that is simple enough to be taught in introductory courses. In [Section 3](#), we provide the idea of the proof of the main result, with technical arguments appearing in the Appendix. We then illustrate the importance of heterogeneity in [Section 4](#), where we show that surplus is not maximal in an LRCE and describe a tax policy that achieves the maximal surplus, and we conclude in [Section 5](#).

see [Clementi and Palazzo \(2016\)](#). There is also a large literature, beginning with the work of [Ericson and Pakes \(1995\)](#), that studies dynamic equilibrium with capital accumulation, stochastic shocks, heterogeneous firms, and entry and exit under *imperfect* competition.

2 The model and main result

2.1 Setup

We adopt Hopenhayn's (1992) infinite-horizon model of a competitive industry. There is an industry with a continuum of potential firms, each of which can produce a homogenous product using production function $f(l, \theta)$, where $l \geq 0$ is an input of production and $\theta \in \Theta = [\theta_L, \theta_H] \subset \mathbb{R}$ is the firm's type. A firm also incurs a fixed cost of production, $FC(\theta)$, which is sunk once the firm decides to enter or stay in the industry and may depend on the firm's type.

Each period $t = 1, 2, \dots$, demand for the product is given by $Q^d(p)$, where $p \geq 0$ is the output price, and the input price is given by inverse input supply function $W(\cdot)$ evaluated at the quantity demanded for the input by the entire industry. In addition, there is an infinite mass of potential entrants with discount factor $\delta \in [0, 1]$ that can decide to enter the market and become a firm. A potential entrant does not know her type, but knows that her type is independently distributed according to the probability measure $\nu \in \Delta(\Theta)$. A firm entering the market must pay a one-time entry cost of $\kappa \geq 0$. After paying the entry cost, a firm immediately learns its own type. Thereafter, types evolve independently across firms according to the conditional probability measure $F(\cdot | \theta) \in \Delta(\Theta)$, where θ is the current type and, for any Borel set $A \subseteq \Theta$, $F(A | \theta)$ is the probability that the type next period will be in A given that the type today is θ . At the end of the period, each firm decides to remain or not in the market for the following period knowing their current type but not knowing their realization of next period's type. There is also an exogenous probability ρ of exit. A firm that exits the market (either endogenously or exogenously) does so permanently and obtains a payoff of zero.

We make the following assumptions on the primitives.

Assumption 1. (i) *There exists $v \in \mathbb{R}_+ \cup \{\infty\}$ such that $Q^d(\cdot)$ is continuously differentiable and strictly decreasing for all $p \in (0, v)$, and $Q^d(p) = 0$ for all $p \geq v$; (ii) $W(\cdot)$ is continuous, nondecreasing and satisfies $W(L) > 0$ for all $L \geq 0$.*

Assumption 2. (i) *f and FC are twice-continuously differentiable; for all $\theta \in \Theta$: $f(\cdot, \theta)$ is increasing, strictly concave, and satisfies $\lim_{l \rightarrow \infty} f(l, \theta) = 0$ and $f(0, \theta) = 0$,*

and $\min_{\theta \in \Theta} \frac{df(0, \theta)}{dl} > 0$; (ii) for all $l \geq 0$, $f(l, \cdot)$ is nonincreasing and $FC(\cdot)$ is non-decreasing, with one of them being strict (i.e., decreasing or increasing, respectively); also, $\nu(\{\theta \in \Theta : FC(\theta) > 0\}) > 0$.

Assumption 3. (i) F is twice-continuously differentiable; (ii) For any $\theta_1 < \theta_2$, $F(\cdot \mid \theta_1)$ first order dominates $F(\cdot \mid \theta_2)$.

Assumption 4. The exogenous probability of exit is positive, i.e., $\rho > 0$.

Assumption 5. (i) ν has a continuous probability density function (pdf), $f_\nu(\cdot)$, such that $\text{supp}(f_\nu) = \Theta$; (ii) For all θ : $F(\cdot \mid \theta)$ has a pdf $f(\cdot \mid \theta)$ and $(\theta', \theta) \mapsto f(\theta' \mid \theta)$ is continuous.

Some of the above assumptions are technical and facilitate the analysis, while others have economic substance. Assumption 1 implies the existence of a downward sloping inverse demand function that we denote by $P^d(\cdot)$; note also that v represents the consumers' maximum willingness to pay. The assumption also requires input prices to be either constant or determined by an upward sloping supply function. Assumption 2 says that production exhibits decreasing returns to scale and that higher types have lower productivity and higher fixed costs. Assumption 3 says that higher types today are more likely to become higher types tomorrow. Together, these assumptions imply that the exit condition is characterized by a threshold.

Assumption 4 guarantees that the life span of a firm is almost surely finite; in particular, if there is no entry, then there must be zero aggregate production in equilibrium. This assumption is made for simplicity and puts the focus on equilibria with positive entry.³

We denote the cost of producing q units for a firm of type θ that faces input price w by

$$C(q, w, \theta) \equiv wf^{-1}(q, \theta) + FC(\theta),$$

³Hopenhayn (1992) instead assumes that $\rho = 0$ and guarantees finite lifespan with an additional recurrence condition on F . He then restricts attention to equilibria with positive entry. His set of equilibria with positive entry and $\rho = 0$ can be viewed in our context as the limit of a sequence of equilibria where ρ goes to zero.

where $f^{-1}(\cdot, \theta)$ is the inverse of the production function of type θ . In the standard textbook analysis, $C(\cdot)$ is usually a primitive of the problem and the wage is fixed and does not vary with aggregate input demand.

Firms take prices as given, and so a firm of type θ faced with prices p and w solves

$$q(p, w, \theta) \equiv \arg \max_q pq - C(q, w, \theta).$$

The corresponding profit is

$$\pi(p, w, \theta) \equiv pq(p, w, \theta) - C(q(p, w, \theta), \theta),$$

and the corresponding input demand is $l(p, w, \theta) \equiv f^{-1}(q(p, w, \theta), \theta)$. Lemma 6 in the appendix lists standard properties of these functions that follow immediately from the previous assumptions.

The expected net present discounted value of a firm that decides to enter or stay in the market, finds out its type is θ , and faces prices p and w every period is

$$V(p, w, \theta) = \pi(p, w, \theta) + \delta(1 - \rho) \max \left\{ \int_{\Theta} V(p, w, \theta') F(d\theta' | \theta), 0 \right\}. \quad (1)$$

We show in the appendix (Lemma 6) that $\pi(p, w, \theta)$ is decreasing in θ , and, consequently, that $V(p, w, \theta)$ is also decreasing in θ . Therefore, the optimal exit decision of firms that face prices p and w every period is characterized by a threshold type. In a steady-state, there will be a marginal type $m \in \Theta$ with the property that all lower types stay and all higher types exit the market.

Let $\mu(n, m)$ denote the steady-state measure of types of firms given the mass of entrants n and the marginal type m . In particular, for any Borel set $A \subseteq \Theta$,

$$\mu(n, m)(A) = \nu(A)n + (1 - \rho) \int_{\theta_L}^m F(A | \theta) \mu(n, m)(d\theta). \quad (2)$$

The assumption that $\rho > 0$ guarantees existence of a steady-state measure.

The corresponding aggregate output supply at prices p and w is

$$Q^s(p, w; n, m) \equiv \int_{\Theta} q(p, w, \theta) \mu(n, m)(d\theta),$$

and the corresponding input demand is

$$L^d(p, w; n, m) \equiv \int_{\Theta} l(p, w, \theta) \mu(n, m)(d\theta).$$

We make two additional assumptions.

Assumption 6. *At least one of the following conditions holds: (i) $W(\cdot)$ is a constant function; or (ii) There exist functions $h(\cdot)$ and $g(\cdot)$ such that $\pi(p, w, \theta) = h(\theta)g(p, w) - FC(\theta)$ for all p, w, θ , where g is continuously differentiable.*

Assumption 7. *If $v < \infty$, then $\pi(v, W(0), \theta_H) > \kappa$.*

Hopenhayn (1992) uses Assumption 6 to show uniqueness of equilibrium, and we will use it to show there is a well-defined equilibrium mapping from aggregate production to the price of the input.⁴ Assumption 7 rules out uninteresting equilibria that have zero aggregate production by requiring that even the highest-cost firm prefers to enter whenever price equals the maximum willingness to pay, v .

Throughout the paper, we maintain assumptions 1-7.

2.2 Long-run competitive equilibrium

Definition 1. A tuple $\langle p^e, w^e, n^e, m^e \rangle$ is a long-run competitive equilibrium (LRCE) if the following conditions are satisfied:

- (i) Output and input market clearing: $Q^d(p^e) = Q^s(p^e, w^e; n^e, m^e)$ and $w^e = W(L^d(p^e, w^e; n^e, m^e))$.
- (ii) Unlimited entry: $\int_{\Theta} V(p^e, w^e, \theta) \nu(d\theta) \leq \kappa$, with equality if $n^e > 0$.
- (iii) Optimal exit: $\int_{\Theta} V(p^e, w^e, \theta') F(d\theta' | m^e) = 0$ if $m^e \in (\theta_L, \theta_H)$, ≥ 0 if $m^e = \theta_H$, and ≤ 0 if $m^e = \theta_L$.

A long-run competitive equilibrium (LRCE) captures the steady state of the industry under the assumption that firms act competitively.⁵ The first condition requires

⁴A sufficient condition for Assumption 6 is provided by Hopenhayn (1992).

⁵Hopenhayn (1992) called an LRCE a stationary equilibrium because, as he showed, it corresponds to the steady state of a perfect foresight equilibrium of the dynamic environment.

market clearing in both the output and input markets and already incorporates the firms' profit maximization condition. The second condition requires the net present value of entry to be equal to the entry cost if there is a positive mass of entrants. It is usually known as the “free entry” condition, but we prefer to use that terminology only for the special case in which entry is indeed free, i.e., $\kappa = 0$. The third condition requires the marginal type to be indifferent between staying or exiting the market, provided it is of course an interior type.

Lemma 1. *In any LRCE, both aggregate production and entry must be positive, i.e., $Q^d(p^e) > 0$ and $n^e > 0$ for any LRCE $\langle p^e, w^e, n^e, m^e \rangle$.*

Proof. Suppose that p^e is an LRCE price such that $Q^d(p^e) = 0$. By Assumption 1, $p^e \geq v$. By profit maximization, it must also be the case that $L^d = 0$, and so $w = W(0)$. By the facts that $\pi(p, W(0), \theta)$ is nondecreasing in p and nonincreasing in θ (see Lemma 6 in the Appendix) and by Assumption 7, $\pi(p^e, W(0), \theta) \geq \pi(v, W(0), \theta) \geq \pi(v, W(0), \theta_H) > \kappa$ for all θ . Thus, $V(p^e, W(0), \theta) > \kappa$ for all θ , so that p^e does not satisfy the entry condition (ii) in Definition 1, thus contradicting that p^e is an LRCE price. In addition, the result that $Q^d(p^e) > 0$ implies, via condition (i) in Definition 1, that $Q^s(p^e, w^e; n^e, m^e) > 0$, which then implies, by the assumption that $\rho > 0$, that $n^e > 0$. \square

2.3 Long-run industry supply function

Definition 2. The long-run industry (inverse) supply function $P_{LR}^s(\cdot)$ is a function $Q \mapsto P_{LR}^s(Q)$ with the property that, for any given $Q > 0$, $p = P_{LR}^s(Q)$ is the unique price that satisfies the following conditions for some $w > 0$, $m \in [\theta_L, \theta_H]$, and $n > 0$:

- (i) $Q = Q^s(p, w; n, m)$ and $w = W(L^d(p, w; n, m))$.
- (ii) $\int_{\Theta} V(p, w, \theta) \nu(d\theta) = \kappa$.
- (iii) $\int_{\Theta} V(p, w, \theta') F(d\theta' \mid m) = 0$ if $m \in (\theta_L, \theta_H)$, ≥ 0 if $m = \theta_H$, and ≤ 0 if $m = \theta_L$.

The (inverse) industry supply function, as envisioned by Marshall and others, maps aggregate production into prices while taking into account the market-clearing conditions in both product and input markets.

The next result follows immediately from the definitions and from Lemma 1, which implies that the entry condition in Definition 2 holds with equality.

Proposition 1. *Suppose that the long-run industry (inverse) supply function $P_{LR}^s(\cdot)$ exists. Then p^e is an LRCE if and only if $p^e = P_{LR}^s(Q^d(p^e))$ and $Q^d(p^e) > 0$.*

Proposition 1 simply says that the LRCE price is such that industry supply equals demand. This result illustrates an appealing feature of the Marshallian approach, which is that we can reduce the equilibrium of the industry to the usual demand-supply diagram.

2.4 Main result

In this section, we provide a characterization of the long-run industry supply function that boils down to the standard textbook treatment in the case where firms are homogeneous. To state the main result, we begin by defining an average aggregate cost function. Letting $\mathcal{M}(\Theta)$ be the space of finite Borel measures that are absolutely continuous with respect to Lebesgue, we define $\bar{C} : [0, \infty)^2 \times \mathcal{M}(\Theta) \rightarrow [0, \infty)$ as

$$\bar{C}(p, w, \eta) = \frac{1}{\eta(\Theta)} \left(\int_{\Theta} C(q(p, w, \theta), w, \theta) \eta(d\theta) + \kappa \right)$$

for all $p \geq 0$, $w > 0$, and $\eta \in \mathcal{M}(\Theta)$. This is the aggregate cost divided by the total mass of firms when the measure of firms is given by $n\eta$, assuming that all firms produce optimally. Note that n enters linearly in the expression for aggregate cost, and so it cancels out when dividing by the total mass. Similarly, let $\bar{q} : [0, \infty)^2 \times \mathcal{M}(\Theta) \rightarrow [0, \infty)$ be the aggregate output divided by the total mass of firms when the measure of firms is given by $n\eta$, defined by

$$\bar{q}(p, w, \eta) = \frac{1}{\eta(\Theta)} \int_{\Theta} q(p, w, \theta) \eta(d\theta).$$

The corresponding *average aggregate cost* is then defined by

$$\bar{AC}(p, w, \eta) \equiv \bar{C}(p, w, \eta) / \bar{q}(p, w, \eta),$$

provided that $\bar{q}(p, w, \eta) > 0$.⁶ Note that the average aggregate cost function is not the weighted average of the average cost function, but rather the total aggregate cost divided by the total aggregate quantity.

⁶If $\bar{q}(p, w, \eta) = 0$, we define $\bar{AC}(p, w, \eta) = \infty$.

Finally, we let $\bar{l} : [0, \infty)^2 \times \mathcal{M}(\Theta) \rightarrow [0, \infty)$ be the weighted input demand, defined by

$$\bar{l}(p, w, \eta) = \frac{1}{\eta(\Theta)} \int_{\Theta} l(p, w, \theta) \eta(d\theta).$$

Next, for each n , m , and δ , we define $\mu_E(n, m, \delta) \in \mathcal{M}(\Theta)$ to be the steady-state measure of types of firms when the mass of entrants is n , the marginal type is $m \in \Theta$, firms exit with exogenous probability $1 - \delta(1 - \rho)$, and the distribution of types for potential entrants is ν , i.e., for any Borel set $A \subseteq \Theta$,

$$\mu_E(n, m, \delta)(A) = \nu(A)n + \delta(1 - \rho) \int_{\theta_L}^m F(A \mid \theta) \mu_E(n, m, \delta)(d\theta).$$

For the special case of $\delta = 1$, $\mu_E(n, m, 1) = \mu(n, m)$ is the *actual* steady-state measure of types defined in equation (2), because in the model firms actually exit with probability ρ , not $1 - \delta(1 - \rho)$.

Finally, it is straightforward to see that μ_E must be linear in n , and so we define

$$\Lambda(m, \delta) \equiv \mu_E(n, m, \delta)/n \in \mathcal{M}(\Theta)$$

to be the measure normalized by the mass of entrants, provided that $n > 0$.

We now state the main characterization result.

Theorem 1. *The long-run industry supply function exists and it is given, for any $Q > 0$, by*

$$P_{LR}^S(Q) = \min_{p, m} \bar{AC}(p, \hat{w}(Q), \Lambda(m, \delta)),$$

where $\hat{w}(Q)$ is the unique solution to

$$w = W \left(Q \frac{\bar{l}(p(w), w, \Lambda(m(w), 1))}{\bar{q}(p(w), w, \Lambda(m(w), 1))} \right) \quad (3)$$

and $\{(p(w), m(w))\} = \arg \min_{p, m} \bar{AC}(p, w, \Lambda(m, \delta))$.

Moreover, the industry supply function $P_{LR}^S(\cdot)$ is continuous, nondecreasing, and it is constant if $W(\cdot)$ is constant.

Theorem 1 extends the standard textbook characterization of the industry supply function to a setting with heterogeneous firms. Consider first the case where $W(\cdot)$

is constant, so that the industry faces a fixed input price irrespective of aggregate production. The industry (inverse) supply function is then horizontal at a price that minimizes the average aggregate cost function, where the minimum is both with respect to price and the marginal type. This result generalizes the standard textbook where all firms are of the same type and the equilibrium price is at the minimum of the average cost function of the representative firm.

If, however, $W(\cdot)$ is increasing, then one needs to take into account that the input price will be affected by aggregate production. For a fixed input wage, price is still given by the minimum of the average aggregate cost. But now the input price depends on aggregate production, via equation (3), and so the price depends on aggregate production. In fact, the relation between price and quantity is monotone, so that the industry supply function is upward sloping.

In particular, Theorem 1 formalizes Marshall's notion of a *representative firm*. Except in the special case where all firms have identical marginal cost functions, however, the average cost function of the representative firm does not generally correspond to the weighted average of the firms' average cost functions. Instead, the relevant average cost function is given by the total aggregate cost divided by total quantity.

Corollary 1. *There exists a unique LRCE and it is characterized by positive entry and positive aggregate production.*

Proof. By Lemma 1 and Proposition 1, we must show that the (inverse) demand and supply functions intersect at a unique point $Q > 0$. By assumption 1, (inverse) demand is decreasing and continuous. By Theorem 1, (inverse) supply is continuous and nondecreasing. Thus, it suffices to show that $\underline{p} \equiv \lim_{Q \rightarrow 0} P_{LR}^S(Q) < v$. Suppose, in order to obtain a contradiction, that $v \leq \underline{p}$. By Theorem 1, there exists m (which one, will not matter) such that $\underline{p} = \bar{AC}(\underline{p}, W(0), \Lambda(m, \delta))$ or, equivalently, $\int \pi(\underline{p}, W(0), \theta) \Lambda(m, \delta)(d\theta) = \kappa$. By Lemma 6(iv) in the Appendix, π is nonincreasing in θ and nondecreasing in v , and so $\pi(v, W(0), \theta_H) \leq \pi(\underline{p}, W(0), \theta_H) \leq \kappa$. But this inequality contradicts assumption 7. \square

Before proving Theorem 1 in Section 3, we provide a simple example.

2.5 A simple example

We discuss an example with three objectives in mind: The example is simple enough to be taught in introductory courses, it conveys much (but not all) of the intuition behind Theorem 1, and it is sufficient to see that standard properties of LRCE with homogeneous firms do not extend to heterogeneous firms.

We assume that: (i) there are only two types, not a continuum, $\theta_H > \theta_L \geq 0$, and each type is equally likely to be drawn by an entrant; (ii) the input price is fixed, and so we omit it from the notation and directly provide as a primitive the cost function; (iii) the cost function takes the particular form $C(q, \theta) = c(q) + \theta$, so that a firm's type represents its fixed cost and all firms have the same marginal cost $MC(q) \equiv c'(q)$; (iv) the entry cost is zero, $\kappa = 0$; (v) types are permanent, so that a firm keeps the type it draws upon entry for its entire lifetime; and (vi) firms are impatient, $\delta < 1$. We assume that the variable cost function $c(\cdot)$ satisfies the following conditions: $c(0) = 0$, $c'(0) = 0$, $c'(q) > 0$ and $c''(q) > 0$ for all $q > 0$, and $\lim_{q \rightarrow \infty} c'(q) = \infty$.

STEADY-STATE MEASURE OF TYPES. It is easy to see that type θ_L will want to stay and type θ_H will exit in equilibrium; in particular, we can drop m from the notation.⁷ The steady-state mass of firms of type θ_L , which we denote by μ_L , is determined by the steady-state mass of entrants, n , as follows:

$$\mu_L = n/2 + \mu_L(1 - \rho). \quad (4)$$

In particular, the mass of firms of type θ_L tomorrow is given by the sum of $n/2$, which is the mass of entrants of type θ_L (recall that half of entrants are of each type), and $\mu_L(1 - \rho)$, which is the mass of firms of type θ_L that were already present and did not exit. The first equation simply says that, in steady state, the mass of firms of type θ_L remains constant. For firms of type θ_H , who never stay for more than one period, their mass is half the mass of entrants. Thus, the masses of firms of each type as a function of the mass of entrants, n , are

$$\mu_L(n) = n/(2\rho) \quad \text{and} \quad \mu_H(n) = n/2. \quad (5)$$

LONG-RUN INDUSTRY SUPPLY FUNCTION. The conditions in the definition of the long-run industry supply function become:

⁷For the free entry condition to hold for $p > 0$, the profit of type θ_H must be negative. Because types are permanent, type θ_H will find it optimal to exit.

- (i) $Q = (\mu_L(n) + \mu_H(n))q(p) > 0$.
- (ii) (Free entry) $NPV(p) \equiv \frac{1}{2}\pi(p, \theta_L)/(1 - \delta(1 - \rho)) + \frac{1}{2}\pi(p, \theta_H) = 0$.

The first condition requires aggregate output supply to equal Q . The free-entry condition requires that the net present value of an entrant is zero. With probability $1/2$, a firm is of type θ_L and remains in the market until it has to exogenously exit, thus expecting a net present value of $\pi(p, \theta_L)/(1 - \delta(1 - \rho))$. With probability $1/2$, a firm is of type θ_H , makes profit $\pi(p, \theta_H)$, and exits the market. In particular, in equilibrium firms make zero profits from an ex ante perspective but make positive profits if they turn out to be a low type and negative profits otherwise.

The weights on the profit functions of each type in the free-entry condition have an intuitive interpretation. Indeed, the weight $\Lambda_L \equiv 1/2(1 - \delta(1 - \rho))$ on $\pi(p, \theta_L)$ is equal to the steady-state mass of firms of type θ_L , normalized by the mass of entrants n , in a hypothetical world where firms, instead of exiting with probability ρ , actually exit with probability $1 - \delta(1 - \rho)$.⁸ In particular, the hypothetical and actual probability of exit coincide as $\delta \rightarrow 1$, and so the weight asymptotically equals the actual steady-state, normalized mass of type θ_L . Similarly, the weight $\Lambda_H \equiv 1/2$ on $\pi(p, \theta_H)$ is equal to the steady-state mass of firms of type θ_H (here, δ is irrelevant because type θ_H exits with probability 1). Thus, the net present value of entry can be written as

$$\begin{aligned} NPV(p) &= \Lambda_L \pi(p, \theta_L) + \Lambda_H \pi(p, \theta_H) \\ &= pq(p)(\Lambda_L + \Lambda_H) - (\Lambda_L C(q(p), \theta_L) + \Lambda_H C(q(p), \theta_H)). \end{aligned} \quad (6)$$

By equation (6), the solution p^e to $NPV(p^e) = 0$ satisfies

$$p^e = AC^e(q(p^e), \Lambda) \equiv \frac{\Lambda_L AC(q(p^e), \theta_L) + \Lambda_H AC(q(p^e), \theta_H)}{(\Lambda_L + \Lambda_H)}, \quad (7)$$

where $\Lambda \equiv (\Lambda_L, \Lambda_H)$, $AC(q, \theta) \equiv C(q, \theta)/q$ is the average cost of type θ , and $q \mapsto AC^e(q, \Lambda)$ is a weighted average cost function. Note that the reason why AC^e is a weighted average of the firms' average cost functions is that each firm has the same marginal cost function, and, therefore, produces the same amount.

By profit maximization, $p^e = MC(q(p^e))$, and so (7) implies that p^e equalizes

⁸Formally, $\Lambda_L \equiv \mu_E(n, \delta)(\theta_L)/n$, where $\mu_E(n, \delta)(\theta_L)$ solves $\mu_E(n, \delta)(\theta_L) = n/2 + \mu_E(n, \delta)(\theta_L)\delta(1 - \rho)$.

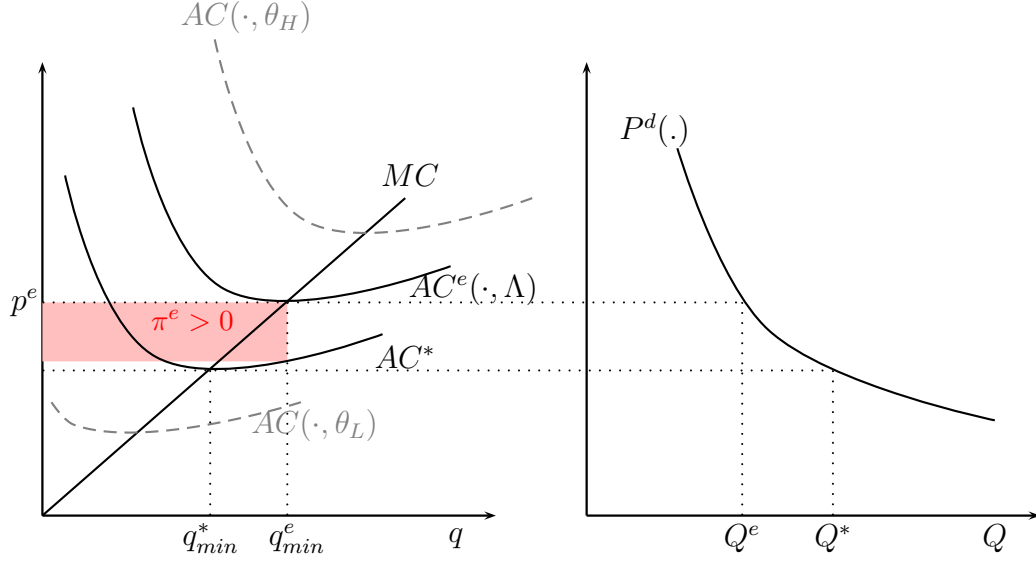


Figure 2: Long-run competitive equilibrium in the example.

marginal and expected average cost,

$$p^e = MC(q(p^e)) = AC^e(q(p^e), \Lambda). \quad (8)$$

The left panel of Figure 2 illustrates how to find the zero-profit price p^e . The figure plots the marginal cost function common to all types, $MC(\cdot)$, the average cost function for each type, $AC(\cdot, \theta)$, and the weighted average cost function $AC^e(\cdot, \Lambda)$. The zero-profit price p^e is given by the intersection of the marginal cost and weighted average cost functions, and this intersection occurs at the minimum point on the weighted average cost function.⁹ Therefore, $q(p^e) = q_{min}^e \equiv \arg \min_q AC^e(q, \Lambda)$ and the zero profit price p^e is given by

$$p^e = AC^e(q_{min}^e, \Lambda) = \min_q AC^e(q, \Lambda).$$

Finally, it is straightforward to check that, given that $p^e > 0$, there exists $n(Q) > 0$ satisfying condition (i) in Definition 2, i.e., $Q = (\mu_L(n(Q)) + \mu_H(n(Q)))q(p^e)$.¹⁰ It then follows that the long-run industry (inverse) supply function exists and is horizontal at the price that minimizes the weighted average cost function $AC^e(\cdot)$. Thus, provided that $P^d(0) > p^e$, there exists a unique LRCE where price is p^e and the mass of entrants

⁹For a proof that the intersection occurs at the minimum point of $AC^e(\cdot, \Lambda)$, note that the first order condition for the problem $\min_q AC^e(q, \Lambda)$ is precisely the condition $MC(q) = AC^e(q, \Lambda)$. Moreover, the second order condition is satisfied because $c''(q) > 0$ for all $q > 0$.

¹⁰In fact, the solution is unique and given by $n(Q) = Q / ((1/2\rho + 1/2)q_{min}^e)$.

n^e is such that the product market clears, i.e., $Q^d(p^e) = (\mu_L(n^e) + \mu_H(n^e))q(p^e)$.¹¹

Figure 2 also illustrates that aggregate profits are strictly positive in an LRCE. The equilibrium profit of the average firm is $\pi^e \equiv (p^e - AC^*(q_{min}^e))q_{min}^e > 0$, where q_{min}^e is the quantity produced in equilibrium, p^e is the price, and

$$AC^*(\cdot) \equiv \frac{(1/(2\rho))AC(\cdot, \theta_L) + (1/2)AC(\cdot, \theta_H)}{((1/(2\rho)) + 1/2)}$$

is the per-unit cost function of the average firm in the industry. The weights in $AC^*(\cdot)$ correspond to the steady-state proportion of firms of each type. As mentioned earlier, these weights converge to Λ as $\delta \rightarrow 1$, but, for $\delta < 1$, $AC^*(\cdot)$ puts more weight on the low cost type relative to $AC^e(\cdot, \Lambda)$. Intuitively, the selection in exit implies that, in the steady-state, the composition of firms is tilted towards low-cost firms relative to the ex-ante perception of a potential entrant who discounts the future. Thus, the fact that potential entrants make zero profits ex-ante implies that the actual firms operating in the steady state make strictly positive profits.

The result that profits are strictly positive in an LRCE changes many of the pre-conceptions of the textbook model of LRCE, such as the idea that 100% of the incidence from tax policy must fall on the demand side if input prices are fixed or that any benefits from a subsidy must accrue exclusively to the owner of a fixed input factor. In Section 4, we focus on another common misconception and show that aggregate surplus is not maximized in an LRCE and that optimal linear taxation would entail taxing profits and subsidizing losses at essentially 100%. The reason for the inefficiency can be visualized in Figure 2, where the steady-state firms are not producing at q_{min}^* , which is the minimum per-unit cost of the average firm that produces in equilibrium, AC^* .

In the special case where there is a single type, $\theta_L = \theta_H$, we obtain the well-known results from the standard textbook analysis: The industry supply function is horizontal at the price that equals the minimum of the average cost function (where all firms have the same cost function), each firm makes zero profits, and aggregate surplus is maximized in an LRCE. Alternatively, we can interpret the standard textbook case as a case where firms are of different types but know their types before entering the market. In that case, only firms of type θ_L will operate in the market in an LRCE.

BEYOND THE SIMPLE EXAMPLE. The main result (Theorem 1) extends the logic

¹¹In fact, the solution is unique and given by $n^e = Q^d(p^e)/((1/2\rho + 1/2)q_{min}^e)$.

in the previous example in several directions. First, marginal cost may differ by types. To see the connection between Theorem 1 and the simple example, note that $\bar{AC}(p, \Lambda) = AC^e(q(p), \Lambda)$, so that the average cost function can always be expressed in terms of price, and we will indeed need to define it in that way for the general case where marginal costs may differ by type. An implication of different marginal costs is that the relevant average cost function will not necessarily be a weighted average of the firms' average cost functions.

Second, the analysis was significantly simplified by assuming that types are fixed. When types follow a Markov process, optimal entry decisions will be characterized as the solution to a dynamic optimization problem. But we can use results from the theory of bounded linear operators to show that ex-ante expected profits can still be expressed as the weighted average of the profits of each type. Third, we will allow for a continuum of firms. By doing so, exit decisions will no longer be trivially characterized and we will see, for example, that exit decisions are also inefficient from the perspective of a planner who wants to maximize aggregate equilibrium surplus. Fourth, Theorem 1 includes the case where the price of the input may vary with aggregate output and the resulting industry supply function may be upward sloping. Finally, the main result also allows for the existence of an entry cost, a straightforward extension that is reflected in the definition of average cost.

3 Proof of Theorem 1

We prove Theorem 1. The first part fixes input price $w > 0$ and shows that there is a unique pair (p, m) that solves conditions (ii)-(iii) in Definition 2 (industry supply function), and that the solution minimizes average aggregate cost. The second part uses this result and condition (i) in Definition 2, to find an expression for input price as a function of aggregate production.

3.1 Part 1. Minimization of \bar{AC}

For each $m \in \Theta$, we first define an operator $\Phi_m : \mathcal{M}(\Theta) \rightarrow \mathcal{M}(\Theta)$ such that, for all $A \subseteq \Theta$ Borel,

$$\Phi_m[\eta](A) = \int_{\theta_L}^m F(A \mid \tilde{\theta}) \eta(d\tilde{\theta}).$$

$\Phi_m[\eta]$ gives the measure of types that results from applying the Markov operator F to those original types that are below the marginal type m , when the measure of original types is given by η .

The next result collects two useful properties of the operator Φ_m .

Lemma 2. (i) For any $\varrho \in [0, 1)$ and $m \in \Theta$, $\sum_{j=0}^{\infty} \varrho^j \Phi_m^j = (I - \varrho \Phi_m)^{-1}$ is a bounded operator from $\mathcal{M}(\Theta)$ to itself, where I is the identity operator; (ii) Let $T_m : L^\infty(\Theta) \rightarrow L^\infty(\Theta)$ be such that, for any $g \in L^\infty(\Theta)$ and $\theta \in \Theta$, $T_m[g](\theta) = 1\{\theta \leq m\} \int_{\Theta} g(\theta') F(d\theta' \mid \theta)$.¹² The adjoint operator of T_m is Φ_m .

Proof. See the Appendix. □

For the proof of Theorem 1, we apply Lemma 2 with $\varrho = \delta(1 - \rho)$. The operator Φ_m can be used to provide an alternative expression for μ_E by noting that

$$\mu_E(n, m, \delta) = \nu n + \varrho \Phi_m[\mu_E(n, m, \delta)].$$

Similarly, we can define $\mu_X(n, m, \delta) \in \mathcal{M}(\Theta)$ as the corresponding measure when the distribution of types for potential entrants is $F(\cdot \mid m)$, i.e.,

$$\mu_X(n, m, \delta) = F(\cdot \mid m)n + \varrho \Phi_m[\mu_X(n, m, \delta)].$$

An implication of Lemma 2(i) is that

$$\Lambda(m, \delta) = \mu_E(n, m, \delta)/n = (I - \delta(1 - \rho)\Phi_m)^{-1}[\nu].$$

Similarly,

$$\Lambda_X(n, m, \delta) \equiv \mu_X(n, m, \delta)/n \equiv (I - \delta(1 - \rho)\Phi_m)^{-1}[F(\cdot \mid m)].$$

A key to the proof of Theorem 1 is to be able to express the value functions in terms of weighted profit functions, where the weights are given by Λ and Λ_X for the entry and exit conditions, respectively. To do this, we define the weighted profit function

¹²As usual, we define $L^\infty(\Theta)$ to be the space of bounded measurable functions.

$\bar{\pi} : [0, \infty)^2 \times \mathcal{M}(\Theta) \rightarrow \mathbb{R}$, where

$$\bar{\pi}(p, w, \eta) = \int \pi(p, w, \theta) \eta(d\theta)$$

for all $p \geq 0$, $w \geq 0$, and $\eta \in \mathcal{M}(\Theta)$. We then state the following two conditions, which the next lemma will link to conditions (ii) and (iii) in Definition 2:

Condition (ii'). $\bar{\pi}(p, w, \Lambda(m, \delta)) = \kappa$.

Condition (iii'). $\bar{\pi}(p, w, \Lambda_X(m, \delta)) = 0$ if $m \in (\theta_L, \theta_H)$, ≥ 0 if $m = \theta_H$, and ≤ 0 if $m = \theta_L$.

Lemma 3. *Fix any $w > 0$. Suppose that (p, m) is the unique solution satisfying conditions (ii') and (iii'). Then (p, m) is also the unique solution satisfying conditions (ii) and condition (iii) in Definition 2.*

Proof. See the Appendix. □

The main insight behind the proof of Lemma 3 is to note that the value function can be written in terms of an operator T_m and that the adjoint of that operator is Φ_m . Consequently, the value function can be written in terms of a weighted average of profits. This is exactly what we did in the example of Section 2.5 (see equation 6). Lemma 3, however, does not yet link LRCE to the minimum of an average cost function, which is the hallmark characterization of LRCE. We now turn to this connection.

Lemma 4. *For any $w > 0$, there is a unique (p, m) satisfying conditions (ii') and (iii'), and it is characterized by*

$$\{(p, m)\} = \arg \min_{p', m'} \bar{AC}(p', w, \Lambda(m', \delta)).$$

Moreover, the solution p satisfies $p = \bar{AC}(p, w, \Lambda(m, \delta)) > 0$ and is continuous and nondecreasing as a function of w .

Proof. See the Appendix. □

We use Figure 3 to describe the intuition behind Lemma 4. In the figure, we plot the characterization of equilibrium from Lemma 3 for a fixed w . The pair

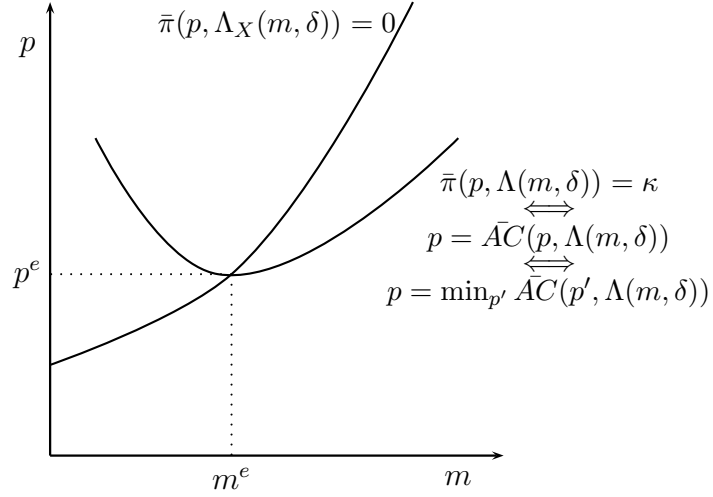


Figure 3: Characterization of entry and exit conditions.

(p^e, m^e) that satisfies conditions (ii) and (iii) in Definition 2 is given by the intersection of the zero entry-profit schedule $\bar{\pi}(p, w, \Lambda(m, \delta)) = \kappa$ and the zero exit-profit schedule $\bar{\pi}(p, w, \Lambda_X(m, \delta)) = 0$ in the (p, m) space. It is not hard to see that the former equation is equivalent to the condition that $p = \bar{AC}(p, w, \Lambda(m, \delta)) = \min_{p'} \bar{AC}(p', w, \Lambda(m, \delta))$. This equivalence is a simple generalization of the standard textbook case; denote the solution of this equation as a function of (m, w) by $\hat{p}(m, w)$. As illustrated by the figure, it is also the case that the zero exit-profit schedule intersects the zero entry-profit schedule at the minimum point of the latter. Thus, m^e minimizes $\bar{AC}(\hat{p}(m, w), \Lambda(m, \delta))$. In other words, (p^e, m^e) jointly minimize \bar{AC} , as stated in Lemma 4.

The reason why the two schedules in Figure 3 intersect at the minimum point of the zero entry-profit schedule is as follows. Consider a point (p, m) where the zero entry-profit schedule lies above the zero exit-profit schedule. At this point, $\bar{\pi}(p, w, \Lambda_X(m, \delta)) > 0$, and so the marginal type m makes a strictly positive profit. If we were to increase the marginal type from m to $m + \varepsilon$, then a potential entrant would now get to stay whenever she draws a type in $(m, m + \varepsilon)$. Since its profit from having a type in the interval would be positive, the firm's ex-ante profit would now increase from zero to a strictly positive number. The price would then need to fall in order to remain on the zero entry-profit schedule. Thus, the zero entry-profit schedule is decreasing whenever it is above the zero exit-profit schedule. By a similar argument, the zero entry-profit is increasing whenever it is below the zero exit-profit schedule.

3.2 Part 2. Clearing of input market

For any $w > 0$, Lemmas 3 and 4 imply that there is a unique solution $p(w)$ and $m(w)$ that satisfies conditions (ii) and (iii) in Definition 2, where, in particular,

$$p(w) = \min_{p', m'} \bar{AC}(p, w, \Lambda(m, \delta)) > 0$$

and $\omega \mapsto p(w)$ is nondecreasing. The final step is to use condition (i) in Definition 2 to express w as a function of Q .

Using $p(w)$ and $m(w)$, condition (i) in Definition 2 can be written as $Q = Q^s(p(w), w; n, m(w))$ and $w = W(L^d(p(w), w; n, m(w)))$, or, equivalently, given the condition in Definition 2 that $n > 0$, $Q = n\bar{q}(p(w), w, \Lambda(m(w), 1))$ and $w = W(n\bar{l}(p(w), w, \Lambda(m(w), 1)))$. Since $Q > 0$, we can solve for $n = Q/\bar{q}(p(w), w, \Lambda(m(w), 1)) > 0$ and replace it into the second equation to obtain

$$w = W\left(Q \frac{\bar{l}(p(w), w, \Lambda(m(w), 1))}{\bar{q}(p(w), w, \Lambda(m(w), 1))}\right). \quad (9)$$

Lemma 5. *For any $Q > 0$, there is a unique solution $\hat{w}(Q)$ to equation (9). Moreover, $Q \mapsto \hat{w}(Q)$ is continuous, nondecreasing, and it is constant if and only if $W(\cdot)$ is constant.*

Proof. See the Appendix. □

The key to Lemma 5 is to use Assumption 6 to show that $f(l(p, w, \theta), \theta)/l(p, w, \theta)$ does not depend on θ and, therefore, \bar{l}/\bar{q} does not depend on Λ . Moreover, \bar{l}/\bar{q} is nondecreasing in p/w due to decreasing returns to scale. These results and the fact that $w/p(w)$ is nondecreasing in w implies the desired result.

The proof of Theorem 1 concludes by noting that the long-run industry supply function is given by $p(\hat{w}(Q))$. In particular, it is continuous in Q (because both $p(\cdot)$ and $\hat{w}(\cdot)$ are continuous), and it is also nondecreasing in Q and constant if $W(\cdot)$ is constant (because $p(\cdot)$ is nondecreasing and $\hat{w}(\cdot)$ is nondecreasing, and constant if $W(\cdot)$ is constant.)

4 Maximal vs. equilibrium surplus

In this section, we show that aggregate surplus is not in general maximized in an LRCE. We then establish that the highest aggregate equilibrium surplus can essentially be achieved by taxing all profits and subsidizing all losses. Moreover, this is the only policy that achieves this level of surplus if the planner is restricted to linear taxes. For simplicity, we assume that the input price is fixed and subsequently omit it from the notation.

4.1 Planner's problem, solution, and comparison to LRCE

An *allocation* is a tuple $\langle Q, n, m, q_A(\cdot) \rangle$, where $Q \geq 0$ is aggregate quantity consumed, $n \geq 0$ is the mass of entrants, $m \in \Theta$ is the marginal type, and $q_A \in L^\infty(\Theta)$ is a function that specifies the quantity produced by each type of firm. An allocation $\langle Q, n, m, q_A(\cdot) \rangle$ is *feasible* if

$$\int_{\Theta} q_A(\theta) \mu(d\theta; n, m) \geq Q. \quad (10)$$

The planner's objective is to choose an allocation $\langle Q, n, m, q_A(\cdot) \rangle$ to maximize steady-state surplus

$$\int_0^Q P^d(\tilde{Q}) d\tilde{Q} - \int_{\Theta} C(q_A(\theta), \theta) \mu(d\theta; n, m) - \kappa n$$

subject to the feasibility constraint in display (10). Steady-state surplus is the sum of consumer's surplus minus the costs of production and entry.

For each tuple $\langle p, n, m \rangle$, we can associate an *induced allocation* $\langle Q^d(p), n, m, q(p, \cdot) \rangle$. The next result characterizes the solution to the planner's problem as the induced allocation for a specific tuple $\langle p^*, n^*, m^* \rangle$.

Proposition 2. *There is a unique allocation that maximizes steady-state surplus; it is given by the allocation induced by $\langle p^*, n^*, m^* \rangle$, where $\langle p^*, n^*, m^* \rangle$ is the unique solution to $Q^d(p^*) = Q^s(p^*; n^*, m^*)$ and*

$$(p^*, m^*) = \arg \min_{p, m} \bar{AC}(p, \Lambda(m, 1)).$$

Proof. See the Appendix. □

Remark. The allocation that maximizes steady-state surplus coincides with the LRCE allocation for the case where $\delta = 1$.

We illustrate the intuition behind Proposition 2 using the example from Section 2.5.

Example, continued. We solve the planner's problem in three steps.

Step 1. Choice of (q_L, q_H) . For a fixed $Q > 0$ and $n > 0$, the planner solves

$$\min_{q_L, q_H} (\mu_L(n) + \mu_H(n)) \left(\frac{\mu_L(n)}{\mu_L(n) + \mu_H(n)} C(q_L, \theta_L) + \frac{\mu_H(n)}{\mu_L(n) + \mu_H(n)} C(q_H, \theta_H) \right)$$

subject to the constraint $\mu_L(n)q_L + \mu_H(n)q_H \geq Q$. The constraint says that total steady-state output must equal Q . The objective function is the steady-state cost of production, written to emphasize the weights of each cost function. The weights correspond to the steady-state probabilities that a firm is of each type, because the planner only cares about the steady-state surplus.¹³

The solution to this problem is to equalize marginal costs, $c'(q_L) = c'(q_H)$, leading to the optimal solution $q_L = q_H = q^*(Q, n) \equiv Q/(\mu_L(n) + \mu_H(n))$. Replacing this solution in the objective function, we obtain the minimized cost

$$(\mu_L(n) + \mu_H(n))C^*(q^*(Q, n)), \tag{11}$$

where

$$\begin{aligned} C^*(q) &= \frac{\mu_L(n)}{\mu_L(n) + \mu_H(n)} C(q, \theta_L) + \frac{\mu_H(n)}{\mu_L(n) + \mu_H(n)} C(q, \theta_H) \\ &= \frac{1}{\rho + 1} C(q, \theta_L) + \frac{\rho}{\rho + 1} C(q, \theta_H) \end{aligned}$$

is the weighted cost function of the planner. The weights correspond to the steady-state probabilities that a firm is of each type. Intuitively, the planner only cares about the steady-state composition of firms.

Step 2. Choice of n . For fixed $Q > 0$ and optimal quantities found in Step 1, the planner chooses n to minimize the cost given by expression (11). This expression is strictly concave in n , and so the following first order condition suffices to characterize

¹³The weights correspond to what we represent by $\Lambda(m, 1)$ in Proposition 2, except that in this simple example there is no marginal type m because we only have two types.

the optimal $n > 0$:

$$AC^*(q^*(Q, n)) \equiv \frac{C^*(q^*(Q, n))}{q^*(Q, n)} = MC(q^*(Q, n)).$$

It is easy to see that the weighted average cost AC^* and the marginal cost MC are equal at the minimum point on the weighted average cost curve. Thus, the optimal mass of entry, n^* , solves $q^*(Q, n^*) = q_{min}^* \equiv \arg \min_q AC^*(q)$. Replacing this solution in the cost function in display (11) yields a minimized cost of $QAC^*(q_{min}^*)$. Intuitively, the planner minimizes cost by producing units at their minimum average cost of production, where the average cost takes into account the steady-state composition of firms.

Step 3. Choice of Q . The final step is to find the aggregate amount of production Q that maximizes total surplus,

$$\max_Q \int_0^Q P^d(\tilde{Q}) d\tilde{Q} - QAC^*(q_{min}^*).$$

The optimal aggregate quantity Q^* is the unique solution to

$$P^d(Q^*) = AC^*(q_{min}^*) = \min_q AC^*(q). \quad (12)$$

Intuitively, the LHS in equation (12) is the consumer's willingness to pay for an extra unit while the RHS is the minimum weighted average cost, which coincides with the marginal cost of producing an extra unit. \square

It follows from Theorems 1 and Proposition 2 that we can compare LRCE with the solution to the planner's problem by simply doing comparative statics with respect to the discount factor. Figure 4 illustrates the main result. As the discount factor increases, both the zero entry-profit and the zero exit-profit schedules characterized in Lemma 3 move to the right. The former does so strictly and the latter does so weakly in the special case of fixed types, and strictly in all other cases. Consequently, the planner always wants the price to be lower than the LRCE price, meaning that he wants higher aggregate production. The effect on exit is, however, ambiguous, as it depends on how much each of the schedules shifts relative to the other one. In the particular case of Figure 4, the planner's optimal exit threshold is higher than the LRCE threshold, but it could also go the opposite way. It is always the case, however,

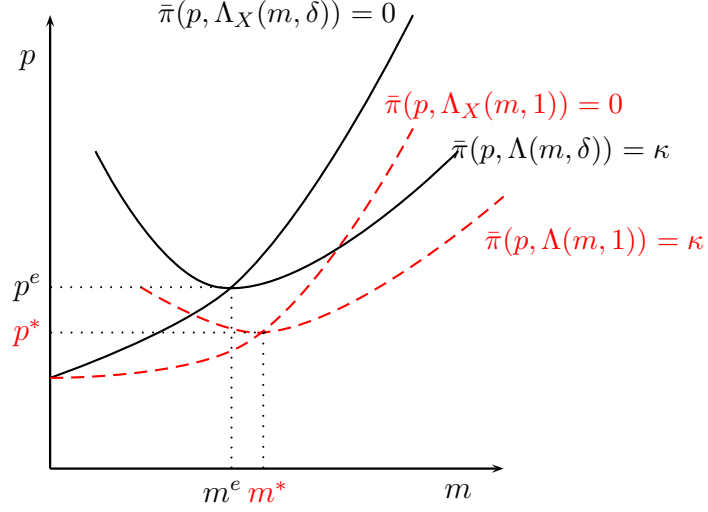


Figure 4: Comparison of equilibrium vs. planner's allocation.

that the planner wants to encourage firms to stay more than desired *at the planner's optimal price* p^* .

Proposition 3. *Suppose that $\delta < 1$. Let $\langle p^e, n^e, m^e \rangle$ denote the LRCE and let $\langle p^*, n^*, m^* \rangle$ be the tuple that induces the planner's optimal solution. It follows that $p^* < p^e$ and, therefore, $Q^d(p^*) > Q^d(p^e)$ and $q(p^*, \cdot) \leq q(p^e, \cdot)$.*

Proof. See the Appendix. □

Example, continued. In the context of the example from Section 2.5, Figure 2 compares the LRCE with the solution to the planner's problem for the case where $\theta_L < \theta_H$ and $\rho < 1$. In particular, the minimum of $AC^*(\cdot)$ is lower than the minimum of $AC^e(\cdot)$, and so the planner prefers a higher aggregate quantity Q^* , a lower quantity per firm q_{min}^* , and a higher mass of entrants n^* compared to the LRCE quantities Q^e , q_{min}^e , and n^e . The reason is that $1/(1 + \rho) > 1/(2 - \delta(1 - \rho))$, and so the function AC^* puts relatively more weight on the low-cost type compared to the function AC^e , implying that $AC^*(q) < AC^e(q)$ for all $q > 0$. Of course, the planner's first-best outcome is not an equilibrium outcome, because the net present value of entry is negative with a mass of entrants higher than the equilibrium mass.

Finally, in the limit, as the discount factor δ goes to 1 and firms become infinitely patient, the planner's solution and the LRCE allocation coincide, i.e., $1/(1 + \rho) = \lim_{\delta \rightarrow 1} 1/(2 - \delta(1 - \rho))$. □

We conclude by noting that there are two simple reasons why the planner's solution differs from the equilibrium outcome. The first feature is that the planner cares about the steady-state outcome, while firms discount the future. This is a typical approach in economics and often adopted either because: (i) the researcher indeed believes that the planner should not discount the future, or (ii) the researcher does not take an explicit stand on the dynamics that lead to equilibrium.¹⁴

But, importantly, the feature that the planner does not discount the future does not, by itself, drive a wedge between the planner's solution and the equilibrium outcome. In fact, no such difference exists if firms are homogeneous, as in the standard textbook case. The second, crucial feature, is that firms are heterogeneous and exit endogenously, and, therefore, selection effects can drive a wedge between the ex-ante and ex-post composition of firms.

4.2 Optimal taxation

We now study how a planner can achieve his first-best outcome via taxation. A tax policy is a tuple $\mathcal{T} = \langle \tau, S^E, S^X \rangle$, where $\tau \in [0, 1]$ is the tax rate on profits (or subsidy if profits are negative), $S^E \in \mathbb{R}$ is a subsidy received for entering, and $S^X \in \mathbb{R}$ is a subsidy received for staying in the industry.

The value function with taxes is now,

$$V_{\mathcal{T}}(p, \theta) = (1 - \tau)\pi(p, \theta) + \delta(1 - \rho) \max \left\{ \int_{\Theta} V_{\mathcal{T}}(p, \theta') F(d\theta' | \theta) + S^X, 0 \right\},$$

where, as usual, p is the equilibrium price and θ is the firm's type.

Definition 3. A tuple $\langle p^e, n^e, m^e \rangle$ is an LRCE with tax policy $\mathcal{T} = \langle \tau, S^E, S^X \rangle$ if the following conditions are satisfied:

- (i) Market clearing: $Q^d(p^e) = Q^s(p^e; n^e, m^e)$.
- (ii) Unlimited entry: $\int V_{\mathcal{T}}(p^e, \theta) \nu(d\theta) + S^E \leq \kappa$, with equality if $n^e > 0$.
- (iii) Optimal exit: $\int_{\Theta} V_{\mathcal{T}}(p^e, \theta') F(d\theta' | m^e) = 0$ if $m^e \in (\theta_L, \theta_H)$, ≥ 0 if $m^e = \theta_H$, and ≤ 0 if $m^e = \theta_L$.

¹⁴Hopenhayn (1992) takes a particular stand on dynamics by focusing on a perfect foresight equilibrium (where an equilibrium entry condition holds every period). He shows that a perfect foresight equilibrium is efficient and that, if it converges, it converges to an LRCE. Thus, under the assumption of perfect foresight and convergence of equilibrium, one could interpret an LRCE as being efficient from the point of view of a planner who discounts the future at the same rate as the firms.

(iv) Self-financing: $\int \tau \pi(p^e, \theta) \mu(n^e, m^e)(d\theta) - (S^E + (\Lambda(m, 1)(\Theta) - 1)S^X)n \geq 0$.

The last condition in Definition 3 requires that, in equilibrium, the planner does not need any outside money to implement the tax policy.

For any tuple $\langle p, n, m \rangle$, let

$$S(p, n, m) \equiv \int_0^{Q^d(p)} P^d(\tilde{Q}) d\tilde{Q} - \int_{\Theta} C(q(p, \theta), \theta) \mu(d\theta; n, m) - \kappa n$$

denote the total surplus of the induced allocation. Because there may exist multiple LRCE under the same tax policy, we define aggregate equilibrium surplus to be the lowest aggregate surplus among all equilibria that share the same tax policy.¹⁵

Definition 4. The *aggregate equilibrium surplus* of tax policy \mathcal{T} is given by $\mathcal{S}(\mathcal{T}) \equiv \inf_{\langle p^e, n^e, m^e \rangle \in \Gamma(\mathcal{T})} S(p, n, m)$, where $\Gamma(\mathcal{T})$ is the set of LRCE with tax policy \mathcal{T} .

The objective of the planner is to choose a tax policy \mathcal{T} to maximize $\mathcal{S}(\mathcal{T})$. By Proposition 2, first-best level of surplus is $\mathcal{S}^* \equiv S(p^*, n^*, m^*)$, where $(p^*, m^*) = \arg \min_{p, m} \bar{AC}(p, \Lambda_{Entry}(m, 1))$ and n^* solves $Q^d(p^*) = Q^s(p^*; n^*, m^*)$. In particular, $\mathcal{S}(\mathcal{T}) \leq \mathcal{S}^*$ for all \mathcal{T} . The next result shows that the planner can approximate \mathcal{S}^* as close as desired using tax policy.

Proposition 4. Suppose that $\delta < 1$. For all $\varepsilon > 0$, there exists a tax policy $\mathcal{T}_\varepsilon = \langle \tau_\varepsilon, S_\varepsilon^E, S_\varepsilon^X \rangle$ such that $\mathcal{S}(\mathcal{T}_\varepsilon) \geq \mathcal{S}^* - \varepsilon$. Moreover, \mathcal{T}_ε satisfies $\lim_{\varepsilon \rightarrow 0} \tau_\varepsilon = 1$, $\lim_{\varepsilon \rightarrow 0} S_\varepsilon^E = \kappa$, and $\lim_{\varepsilon \rightarrow 0} S_\varepsilon^X = 0$.

Proof. See the Appendix. □

We will illustrate the intuition behind Proposition 4 using the example from Section 2.5.

Example, continued. The NPV of entry with taxation is now

$$NPV_{\mathcal{T}}(p) = (1 - \tau)NPV(p) + S^E,$$

¹⁵For example, if $\tau = 1$, $S^E = \kappa$, and $S^X = 0$, firms are indifferent about all decisions and so there exist multiple equilibria.

where $NPV(\cdot)$ is the NPV without taxation introduced in Section 2.5 and S^E is the entry subsidy. The self-financing constraint becomes

$$S^E \leq \tau R(p) \equiv \tau \left(\frac{\mu_L(n)}{n} \pi(p, \theta_H) + \frac{\mu_H(n)}{n} \pi(p, \theta_L) \right). \quad (13)$$

The LHS of equation (13) is the entry subsidy per entrant while the RHS is the tax revenue per entrant at tax rate τ and price p . Note that the RHS does not depend on n .

Letting p^e denote the equilibrium price without taxes and p^* the price that maximizes aggregate surplus, we now show that, for any $\varepsilon \in (0, p^e - p^*]$, the planner can obtain price $\hat{p} = p^* + \varepsilon$ in an LRCE with taxation. Thus, the planner can implement any price in the interval $(p^*, p^e]$ and, in particular, can implement a price as close as desired to its preferred price p^* .

To see this claim, let

$$\tau_\varepsilon = \frac{-NPV(p^* + \varepsilon)}{-NPV(p^* + \varepsilon) + R(p^* + \varepsilon)}$$

and $S_\varepsilon^E = \tau_\varepsilon R(p^* + \varepsilon)$. It follows that

$$NPV_{\tau}(p) = \frac{R(p^* + \varepsilon)}{-NPV(p^* + \varepsilon) + R(p^* + \varepsilon)} (NPV(p) - NPV(p^* + \varepsilon)),$$

which is zero at $p = p^* + \varepsilon$, thus establishing the claim. Note also that, as $\varepsilon \rightarrow 0$, then $S_\varepsilon \rightarrow 0$ (since by definition of p^* , aggregate profits are zero at p^*) and $\tau_\varepsilon \rightarrow 1$, essentially taxing all profits and subsidizing all losses.

The left panel of Figure 5 illustrates the intuition behind this result in two steps. The first step is to notice that a tax on profits by itself does not affect the LRCE price but serves to flatten the NPV curve. The second step is to introduce a small subsidy so that the flattened curve plus the subsidy intersects the horizontal axis at a lower price. To get as close as desired to p^* , the subsidy must be small, because aggregate profits at prices slightly above p^* are small, and so the collected revenue is small. But the planner can choose τ high enough so that a small subsidy is indeed enough to take the taxed NPV curve, with a slope close to zero, to an equilibrium price near p^* .¹⁶ \square

¹⁶The policy $\tau = 1$ and $S = 0$ exactly achieves the first best, but, because firms under this policy are indifferent about entry irrespective of the price, it also achieves several other suboptimal equilibria. Alternatively, it is possible to show that the first-best can be exactly achieved by taxing

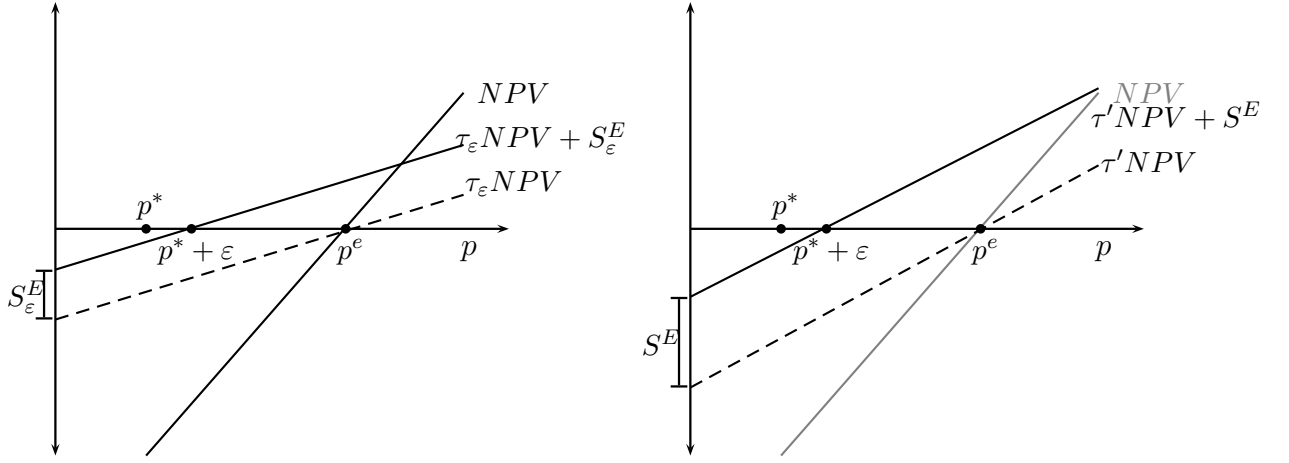


Figure 5: Optimal taxation in the example.

A natural question is whether there is an alternative tax policy that balances the budget, arbitrarily approximates the first-best outcome, but does not tax profits at essentially 100 percent. The next results says the answer is no.

Proposition 5. *Suppose that $\delta < 1$. Consider any sequence of tax policies $(\mathcal{T}_j = \langle \tau_j, S_j^E, S_j^X \rangle)_j$ with corresponding sequence of equilibria $\langle p_j^e, n_j^e, m_j^e \rangle$ such that $\lim_{j \rightarrow \infty} S(p_j^e, n_j^e, m_j^e) = \mathcal{S}^*$. Then $\lim_{j \rightarrow \infty} \tau_j = 1$.*

Proof. See the Appendix. □

Example, continued. The right panel of Figure 5 illustrates this result for the example of Section 2.5. Consider, for example, the NPV schedule associated with a tax rate τ' and zero subsidy depicted in the figure. The subsidy necessary to make sure that this schedule with the added subsidy intersects the horizontal axis near p^* is substantially greater than zero, but the balanced budget condition restricts the subsidy to be very close to zero because $R(p^*) = 0$. □

only positive, but not negative, profits at 100% and providing a positive entry subsidy to compensate for the possibility that an entrant draws a high type. This approach, however, does not extend to the general case in Proposition 4, where the planner may also want to affect exit decisions.

5 Conclusion

One of Marshall's pathbreaking contributions was to envision a model of long-run competitive equilibrium (LRCE) where price is determined by the characteristics of a representative firm. But, except in the case where all firms in the industry are identical, the identity of this representative firm is unclear. Using Hopenhayn's (1992) model, we extend the classical theory of LRCE to the case of heterogeneous firms and show that, under certain conditions, the (long-run) industry supply function with heterogeneous firms exists and can indeed be characterized as the solution to the minimization problem of a "representative" average cost function. The representative average cost function envisioned by Marshall, however, is not generally a weighted average cost function of all producing firms, but rather the aggregate cost divided by the aggregate quantity.

As an application of the importance of accounting for heterogeneity, we show that maximal surplus is not maximized in an LRCE and that the only way to approximate the maximal surplus with a linear tax is to tax all profits and subsidize all losses. This result is absent in discussions about the properties of competitive markets and it is driven both by the focus on long-run equilibrium surplus (which is also the focus in the standard model with homogeneous firms) as well as by selection-in-survival in an industry with heterogeneous firms.

References

- Allen, Roy G. D., London School of Economics, and Political Science,** "Mathematical analysis for economists," Technical Report, Macmillan London 1938.
- Ambrosetti, Antonio and Giovanni Prodi,** *A primer of nonlinear analysis* number 34, Cambridge University Press, 1995.
- Barone, E.,** "On the analysis of dynamic problems," in *Italian Economic Papers*, L. L. Pasinetti, ed., 1992.
- Clementi, Gian Luca and Berardino Palazzo,** "Entry, Exit, Firm Dynamics, and Aggregate Fluctuations," *American Economic Journal: Macroeconomics*, July 2016, 8 (3), 1–41.

- Ericson, Richard and Ariel Pakes**, “Markov-perfect industry dynamics: A framework for empirical work,” *The Review of Economic Studies*, 1995, 62 (1), 53–82.
- Hicks, John R.**, *Value and Capital: An Inquiry Into Some Fundamental Principles of Economic Theory*, Oxford: Clarendon Press, 1946.
- Hopenhayn, Hugo A.**, “Entry, exit, and firm dynamics in long run equilibrium,” *Econometrica*, 1992, pp. 1127–1150.
- Jovanovic, Boyan**, “Selection and the Evolution of Industry,” *Econometrica*, 1982, pp. 649–670.
- Kaldor, Nicholas**, “The equilibrium of the firm,” *The Economic Journal*, 1934, 44 (173), 60–76.
- Lucas, Robert E.**, “Adjustment costs and the theory of supply,” *The Journal of Political Economy*, 1967, pp. 321–334.
- Lucas, Robert E. and Edward C. Prescott**, “Investment under uncertainty,” *Econometrica*, 1971, pp. 659–681.
- Marshall, Alfred**, “Principles of economics: an introductory volume,” 1920.
- Opocher, Arrigo and Ian Steedman**, “The industry supply curve: Two different traditions,” *The European Journal of the History of Economic Thought*, 2008, 15 (2), 247–274.
- Pigou, Arthur C.**, “An analysis of supply,” *The economic journal*, 1928, 38 (150), 238–257.
- Viner, Jacob**, “Cost curves and supply curves,” in *Readings in price theory*, G. J. Stigler and K. E. Boulding, eds., 1953, pp. 198–232.

A Appendix

A.1 Preliminary lemma

Lemma 6. *The following are true:*

(i) C is continuous and twice continuously differentiable; $\frac{dC(q,w,\theta)}{d\theta} > 0$; and, for any $w > 0$, $MC(q,w,\theta) \equiv \frac{dC(q,w,\theta)}{dq} > 0$ and $\frac{d^2C(q,w,\theta)}{dq^2} > 0$.

(ii) For all θ , $q(\cdot, \cdot, \theta)$ is homogeneous of degree 0; For $w > 0$, $(p, \theta) \mapsto q(p, w, \theta)$ is continuously differentiable and $p \mapsto q(p, w, \theta)$ is nondecreasing and increasing for all p such that $q(p, w, \theta) > 0$; For all (p, w, θ) , $q(p, w, \theta) (MC(q(p, w, \theta), w, \theta) - p) = 0$; $\lim_{p \rightarrow \infty} q(p, w, \theta) = \infty$;

(iii) For all θ , $l(\cdot, \cdot, \theta)$ is homogeneous of degree 0; $\frac{p}{w} \mapsto l(p, w, \theta)$ is nondecreasing and strictly increasing for p/w such that $l(p, w, \theta) > 0$; $\lim_{p \rightarrow \infty} l(p, w, \theta) = \infty$.

(iv) π is continuously differentiable; for all θ , $\pi(\cdot, \cdot, \theta)$ is homogeneous of degree 0; $\theta \mapsto \pi(p, w, \theta)$ is decreasing and $p \mapsto \pi(p, w, \theta)$ is nondecreasing and increasing when $l(p, w, \theta) > 0$; $w \mapsto \pi(p, w, \theta)$ is nonincreasing and decreasing when $l(p, w, \theta) > 0$; for any $w > 0$, $\liminf_{p \rightarrow \infty} \inf_{\theta \in \Theta} \pi(p, w, \theta) = \infty$.

(v) For any $p \geq 0$ and $w > 0$: $V(p, w, \cdot)$ is unique and belongs to $\mathbb{D}(\Theta)$; there exists a threshold $m \equiv m(p, w)$ such that $M[V(p, w, \cdot)](\theta) < (>) 0$ iff $\theta > (<) m$; $p \mapsto V(p, w, \theta)$ is nondecreasing (increasing when $l(p, w, \theta) > 0$) and $w \mapsto V(p, w, \theta)$ is nonincreasing (decreasing when $l(p, w, \theta) > 0$); $(\theta, p) \mapsto V(p, w, \theta)$ is continuous; For all θ , $V(\cdot, \cdot, \theta)$ is homogeneous of degree 0.

(vi) Let V_ϱ denote the value function define in equation (1), where we now make the dependence of V on $\varrho \equiv (1 - \rho)\delta \in [0, 1)$ explicit. Then $\varrho \mapsto V_\varrho(p, w, \theta)$ is nondecreasing.

Proof. (i) By definition of C , $C(q, w, \theta) = wf^{-1}(q, \theta) + FC(\theta)$. By assumption 2(i), it is easy to see that C is continuous and twice continuously differentiable. Also, $\frac{dC(q,w,\theta)}{dq} = w \frac{df^{-1}(q,\theta)}{dq}$ and $\frac{d^2C(q,w,\theta)}{dq^2} = w \frac{d^2f^{-1}(q,\theta)}{dq^2}$. By assumption 2(i), $\frac{df^{-1}(q,\theta)}{dq} > 0$ and $\frac{d^2f^{-1}(q,\theta)}{dq^2} < 0$, so, when $w > 0$, $\frac{dC(q,w,\theta)}{dq} > 0$ and $\frac{d^2C(q,w,\theta)}{dq^2} < 0$. Finally, $\frac{dC(q,w,\theta)}{d\theta} = w \frac{df^{-1}(q,\theta)}{d\theta} + \frac{dFC(\theta)}{d\theta}$, so $\frac{dC(q,w,\theta)}{d\theta} > 0$ follows from assumption 2(ii).

(ii) By part (i), under $w > 0$, the FOC for q (which is also sufficient) is given by, $q(p, w, \theta) (MC(q(p, w, \theta), w, \theta) - p) = 0$ for any p and $\theta \in \Theta$, where $MC(q, w, \theta) = w \frac{df^{-1}(q,\theta)}{dq}$. Let $q^*(p, w, \theta)$ be the solution to $p = w \frac{df^{-1}(q,\theta)}{dq}$. Clearly, $q^*(p, w, \theta) = q^*(\lambda p, \lambda w, \theta)$ for any $\lambda > 0$. Since $q(p, w, \theta) = q^*(p, w, \theta)$ for (p, w, θ) such that $q(p, w, \theta) > 0$ and equal to zero otherwise, then $q(p, w, \theta) = q(\lambda p, \lambda w, \theta)$ for any $\lambda > 0$. By part (i) and the implicit function theorem, $(p, \theta) \mapsto q(p, \theta)$ is continuously

differentiable with (partial) derivatives given by

$$\frac{dq(p, w, \theta)}{d\theta} = - \frac{q(p, w, \theta) \frac{dMC(q(p, w, \theta), w, \theta)}{d\theta}}{\left(MC(q(p, \theta), w, \theta) - p + q(p, w, \theta) \frac{d^2C(q(p, \theta), w, \theta)}{dq^2} \right)}$$

and

$$\frac{dq(p, w, \theta)}{dp} = \frac{q(p, w, \theta)}{\left(MC(q(p, w, \theta), w, \theta) - p + q(p, w, \theta) \frac{d^2C(q(p, w, \theta), w, \theta)}{dq^2} \right)}.$$

The last display and the fact that $MC(q(p, w, \theta), w, \theta) - p \geq 0$ and $\frac{d^2C(q(p, w, \theta), w, \theta)}{dq^2} > 0$ shows that $p \mapsto q(p, w, \theta)$ is nondecreasing, and increasing if $q(p, w, \theta) > 0$.

We now show that $\lim_{p \rightarrow \infty} q(p, w, \theta) = \infty$. Since $p = w \frac{df^{-1}(q(p, w, \theta), \theta)}{dq}$, it follows that, as $p \rightarrow \infty$, $\frac{df^{-1}(q(p, w, \theta), \theta)}{dq} \rightarrow \infty$. Since $\frac{df^{-1}(q, \theta)}{dq} = \frac{1}{df(f^{-1}(q, \theta), \theta)/dl}$, it follows that, under assumption 2, $f^{-1}(q(p, w, \theta), \theta) \rightarrow \infty$ or, equivalently, $\lim_{p \rightarrow \infty} q(p, w, \theta) = \infty$.

(iii) Homogeneity of $l(., ., \theta)$ follows from homogeneity of $q(., ., \theta)$ and the fact that $l(p, w, \theta) = f^{-1}(q(p, w, \theta), \theta)$. It follows from part (ii) that $\frac{p}{w} \mapsto q(p, w, \theta)$ is nondecreasing and strictly increasing for p/w such that $q(p, w, \theta) > 0$. Therefore, so is $l(p, w, \theta)$. Similarly, the fact that $\lim_{p \rightarrow \infty} q(p, w, \theta) = \infty$ implies that $\lim_{p \rightarrow \infty} l(p, w, \theta) = \infty$.

(iv) Homogeneity of $\pi(., ., \theta)$ follows from homogeneity of $q(., ., \theta)$ and straightforward calculations. It is also easy to see that $(p, w, \theta) \mapsto pq - C(q, w, \theta)$ is continuous. Thus by the Theorem of the Maximum, π is continuous. Take any $\theta_1 < \theta_2$. By optimality $\pi(p, w, \theta_1) \geq pq(p, w, \theta_2) - C(q(p, w, \theta_2), w, \theta_1)$ and by the fact that $\frac{dC(q, w, \theta)}{d\theta} > 0$ (by part (i)), $\pi(p, w, \theta_1) > \pi(p, w, \theta_2)$. Similarly, take any $p_1 < p_2$. By optimality $\pi(p_2, w, \theta) \geq p_2q(p_1, w, \theta) - C(q(p_1, w, \theta), w, \theta) \geq \pi(p_1, \theta)$. It follows that $\pi(p, w, \theta) = pq(p, w, \theta) - wf^{-1}(q(p, w, \theta), \theta) - FC(\theta)$. By the envelope theorem $\frac{d\pi(p, w, \theta)}{dw} = -f^{-1}(q(p, w, \theta), \theta) = -l(p, w, \theta) \leq 0$ and with strict inequality if $l(p, w, \theta) > 0$.

For any $w > 0$, we show that there exists a $p \equiv p(w)$ such that $\inf_{\theta \in \Theta} q(p, w, \theta) > 0$. If this exist, then $\inf_{\theta \in \Theta} \pi(p', w, \theta) \geq p' \inf_{\theta \in \Theta} q(p, w, \theta) - \sup_{\theta \in \Theta} C(q(p, w, \theta), \theta)$ for any $p' \geq 0$. By parts (i)-(ii) $\theta \mapsto C(q(p, w, \theta), \theta)$ is continuous (and thus bounded in Θ), so $\sup_{\theta \in \Theta} C(q(p, w, \theta), \theta) < \infty$ and thus, by taking $p' \rightarrow \infty$, it follows that $\liminf_{p \rightarrow \infty} \inf_{\theta \in \Theta} \pi(p', w, \theta) = \infty$. To show the existence of such $p(w)$, suppose it does not exist. Then there exists a $w > 0$ and a sequence of $(\theta(p))_p$ such that $p = \frac{df^{-1}(0, \theta(p))}{dq} = \frac{1}{df(0, \theta(p))/dl}$ (note that $f^{-1}(0, \theta) = 0$) for any p . Taking $p \rightarrow \infty$, it follows

that $df(0, \theta(p))/dl \rightarrow 0$, but this violates assumption the fact that $\inf_{\theta \in \Theta} \frac{df(0, \theta)}{dl} > 0$ (under assumption 2(i)).

(v) Let $\varrho \equiv (1 - \rho)\delta$ and let $\mathbb{D}(\Theta)$ the space of continuous decreasing functions in Θ . (i) V can be cast as $V(p, w, \theta) = \pi(p, w, \theta) + \varrho \max \{M[V(p, w, \cdot)](\theta), 0\}$. By part (iv), $\pi(p, w, \cdot) \in \mathbb{D}(\Theta)$, so, under assumptions 3(ii) and 5 (which implies that M maps the space of continuous function into itself), by standard arguments, the Bellman operator $g \mapsto B_\delta[g](\cdot) \equiv \pi(p, w, \cdot) + \varrho \max \{M[g](\cdot), 0\}$ maps $\mathbb{D}(\Theta)$ into itself. Since $\varrho \in [0, 1)$, by standard arguments one can show that B_ϱ satisfies Blackwell's sufficient conditions and thus it is a contraction and thus there exists a unique $V(p, w, \cdot) \in \mathbb{D}(\Theta)$. Note that $V(p, w, \cdot) \in \mathbb{D}(\Theta)$ and also that for each $g \in \mathbb{D}(\Theta)$, there exists a threshold m_g such that $M[g](\theta) < (>)0$ iff $\theta > (<)m_g$. Thus, letting $m(p, w) \equiv m_{V(p, w, \cdot)}$ the result follows. By part (iv) $p \mapsto \pi(p, w, \theta)$ is nondecreasing, so by standard arguments, $p \mapsto V(p, w, \theta)$ is nondecreasing. Similarly, by part (iv) $w \mapsto \pi(p, w, \theta)$ is nonincreasing, the same type of arguments yield that $w \mapsto V(p, w, \theta)$ is nonincreasing. Homogeneity of $V(\cdot, \cdot, \theta)$ follows from homogeneity of $\pi(\cdot, \cdot, \theta)$ and standard calculations. Continuity of $(\theta, p) \mapsto V(p, w, \theta)$ follows from the fact that π is continuous and $(p, \theta) \mapsto M[V(p, w, \cdot)](\theta)$ is too.

(vi) We already established in part (v) that B_ϱ is a contraction. Since $\max \{M[g](\cdot), 0\} \geq 0$ for any g and any $V \in \mathbb{D}(\Theta)$, it follows that $B_{\varrho_1}[V] \geq B_{\varrho_2}[V]$ for any $\varrho_1 \geq \varrho_2$. Applying this to $V = V_{\varrho_2}$ and noting that $B_{\varrho_2}[V_{\varrho_2}] = V_{\varrho_2}$, by monotonicity of B_{ϱ_1} , it follows that $B_{\varrho_1}^2[V_{\varrho_2}] \geq B_{\varrho_1}[B_{\varrho_2}[V_{\varrho_2}]] \geq B_{\varrho_1}[V_{\varrho_2}] \geq B_{\varrho_2}[V_{\varrho_2}] = V_{\varrho_2}$. Iterating in this fashion and noting that B_{ϱ_1} is a contraction, it follows that $V_{\varrho_1} = \lim_{t \rightarrow \infty} B_{\varrho_1}^t[V_{\varrho_2}] \geq V_{\varrho_2}$. \square

A.2 Proof of Lemma 2

Proof. [Proof of Lemma 2] Let $L(\mathcal{M}(\Theta))$ denote the space of linear bounded operators mapping $\mathcal{M}(\Theta)$ to itself. (i) By assumption in the Lemma, $\varrho \|\Phi_m\| < 1$ (here $\|\cdot\|$ is the operator norm¹⁷), it is easy to see that the sequence of partial sums $(\sum_{j=0}^n \varrho^j \Phi_m^j)_n$ is Cauchy (under the operator norm). By completeness of $L(\mathcal{M}(\Theta))$ it follows that $S \equiv \sum_{j=0}^\infty \varrho^j \Phi_m^j \in L(\mathcal{M}(\Theta))$. It is easy to see that $\varrho \Phi_m S = S - I$ or equivalently $(I - \varrho \Phi_m)S = I$; similarly $S(I - \varrho \Phi_m) = I$. This implies that S is the inverse of $(I - \varrho \Phi_m)$ which is denoted by $(I - \varrho \Phi_m)^{-1}$.

¹⁷The space $\mathcal{M}(\Theta)$ is equipped with the total variation norm and the operator norm $\|\Phi_m\| \equiv \sup_{\eta \neq 0} \frac{\|\Phi_m[\eta]\|_{TV}}{\|\eta\|_{TV}} \leq 1$ where $\|\eta\|_{TV} \equiv 0.5 \int_\Theta |f_\eta(\theta)| d\theta$ where f_η is the Radon-Nikodym derivative of η with respect to Lesbegue.

(ii) For any $g \in L^\infty(\Theta)$ and any η Borel measure of Θ . By Fubini's Theorem,

$$\int_{\Theta} T_m[g](\theta) \eta(d\theta) = \int_{\Theta} g(\theta') \left\{ \int 1\{\theta \leq m\} F(d\theta' \mid \theta) \eta(d\theta) \right\} = \int_{\Theta} g(\theta') \Phi_m[\eta](d\theta').$$

□

A.3 Proof of Lemma 3

Lemma 7. For any $(p, w, \theta) \in \mathbb{R}_+^2 \times \Theta$ and any $m \in \Theta$, let

$$V_m(p, w, \theta) = \pi(p, w, \theta) + \varrho T_m[V_m(p, w, \cdot)](\theta)$$

where $T_m[g](\theta) = 1\{\theta \leq m\} \int_{\Theta} g(\theta') F(d\theta' \mid \theta)$. Then, for any $\lambda \in M(\Theta)$,

$$\int V_m(p, w, \theta) \lambda(d\theta) = \int \pi(p, w, \theta) (I - \varrho \Phi_m)^{-1} [\lambda](d\theta).$$

Proof. T_m maps $\mathbb{B}(\Theta)$ into itself and it is easy to see that $V_m(p, w, \cdot) \in \mathbb{B}(\Theta)$. We can cast V_m as

$$V_m(p, w, \theta) = \pi(p, w, \theta) + \varrho T_m[V_m(p, w, \cdot)](\theta) = \sum_{j=0}^J \varrho^j T_m^j[\pi(p, w, \cdot)](\theta) + \varrho^{J+1} T_m^{J+1}[V_m(p, w, \cdot)](\theta),$$

where $T_m^0 = I$. Since $V(p, w, \cdot) \in \mathbb{B}(\Theta)$, T_m maps $\mathbb{B}(\Theta)$ into itself, and $\varrho \in [0, 1)$, then $\limsup_{J \rightarrow \infty} \varrho^{J+1} T_m^{J+1}[V_m(p, w, \cdot)](\theta) = 0$. Thus, we have established the following equivalent representation of the value function: For any $(p, w, \theta) \in \mathbb{R}_+^2 \times \Theta$ and any $m \in \Theta$,

$$V_m(p, w, \theta) = \sum_{j=0}^{\infty} \varrho^j T_m^j[\pi(p, w, \cdot)](\theta).$$

By Lemma 2(ii), the adjoint operator of T_m is Φ_m . It is easy to see that the adjoint operator of T_m^j is Φ_m^j for any $j \in \{1, 2, \dots\}$ and thus, for any $p \geq 0$

$$\int V_m(p, w, \theta) \lambda(d\theta) = \int \pi(p, w, \theta) \left(\sum_{j=0}^{\infty} \varrho^j \Phi_m^j[\lambda](d\theta) \right).$$

By Lemma 2(i), $\sum_{j=0}^{\infty} \varrho^j \Phi_m^j[\cdot] = (I - \varrho \Phi_m)^{-1}[\cdot]$ and thus

$$\int V_m(p, w, \theta) \lambda(d\theta) = \int \pi(p, w, \theta) (I - \varrho \Phi_m)^{-1}[\lambda](d\theta).$$

□

Proof. [Proof of Lemma 3] Throughout the proof, fix $w > 0$. Also, let $\theta \mapsto M[g](\theta) \equiv \int g(\theta') F(d\theta' | \theta)$ for any $g \in \mathbb{C}(\Theta)$. Let \mathcal{S}' be the set of (p, m) such that conditions (ii') and (iii') are satisfied. Similarly, let \mathcal{S} be the set of (p, m) such that conditions (ii) and condition (iii) in Definition 2 hold.

We first show that $\mathcal{S} \subseteq \mathcal{S}'$. Let $(p, m) \in \mathcal{S}$. By definition of m , $V = V_m$. By invoking Lemma 7, first with $\lambda = \nu$ and then with $\lambda = F(\cdot | m)$, it follows that

$$\kappa = \int V(p, w, \theta) \nu(d\theta) = \int \pi(p, w, \theta) \Lambda(m, \delta)(d\theta) = \bar{\pi}(p, w, \Lambda(m, \delta)),$$

and, if $m \in (\theta_L, \theta_H)$,

$$0 = \int V(p, w, \theta) F(d\theta | m) = \int \pi(p, w, \theta) \Lambda_X(m, \delta)(d\theta) = \bar{\pi}(p, w, \Lambda_X(m, \delta)).$$

(≥ 0 if $m = \theta_H$, and ≤ 0 if $m = \theta_L$). Therefore, $(p, m) \in \mathcal{S}'$.

We now show that if $\mathcal{S}' = \{(p', m')\}$ (i.e., it is a singleton), then \mathcal{S} is nonempty (and equal to $\{(p', m')\}$). By assumption, $\int \pi(p', w, \theta) \Lambda(m', \delta)(d\theta) = \kappa$ and $\int \pi(p', w, \theta) \Lambda_X(m', \delta)(d\theta) = 0$ if $m' \in (\theta_L, \theta_H)$, ≥ 0 if $m' = \theta_H$, and ≤ 0 if $m' = \theta_L$. By Lemma 7 with $\lambda = \nu$ and $\lambda = F(\cdot | m)$, it follows that $\int V_{m'}(p', w, \theta) \nu(d\theta) = \kappa$ and $M[V_{m'}(p', w, \cdot)](m') = 0$ if $m' \in (\theta_L, \theta_H)$, ≥ 0 if $m' = \theta_H$, and ≤ 0 if $m' = \theta_L$. For any (p, w) , let $m(p, w) \in \Theta$ be given by $M[V(p, w, \cdot)](m(p, w)) = 0$ if $m \in (\theta_L, \theta_H)$, ≥ 0 if $m(p, w) = \theta_H$, and ≤ 0 if $m(p, w) = \theta_L$. By construction of V_m it follows that $V_{m(p, w)}(p, w, \cdot) = V(p, w, \cdot)$, so it is straightforward to check that $(p', m(p', w)) \in \mathcal{S}'$, but by unicity, $m' = m(p', w)$. Finally, note that it also holds that $\int V_{m'}(p', w, \theta) \nu(d\theta) = \int V_{m(p', w)}(p', w, \theta) \nu(d\theta) = \int V(p', w, \theta) \nu(d\theta) = \kappa$, therefore $(p', m(p', w)) \in \mathcal{S}$. The first part of the proof guarantees that $\mathcal{S} = \mathcal{S}'$. □

A.4 Proof of Lemma 4

Since any $\eta \in \mathcal{M}(\Theta)$ has a Radon-Nikodym derivative with respect to Lebesgue, f_η , an alternative representation of Φ_m consists of an operator from $a_m : L^1(\Theta) \rightarrow L^1(\Theta)$ given by $a_m[g](\cdot) \equiv \int 1\{\theta \leq m\} f(\cdot \mid \theta) g(\theta) d\theta$. That is $a_m[g]$ is the Radon-Nikodym derivative of $\Phi_m[\mu_g]$ with $\mu_g(A) = \int_A g(\theta) d\theta$. By Assumption 5, $\int \sup_{t \in \Theta} f(t \mid \theta) d\theta < \infty$, and therefore, by the DCT, a_m maps into the space of continuous bounded functions, $\mathbb{C}(\Theta)$.

Lemma 8. *Then, for any $\varrho \in [0, 1]$ and any $m \in \Theta$: (i) $\Lambda(m, \delta)$ has a Radon-Nikodym derivative with respect to Lebesgue, $f_E(m, \delta) \in \mathbb{C}(\Theta)$, and it is given by*

$$f_E(m, \delta) = (I - \varrho a_m)^{-1} [f_\nu];$$

(ii) $\Lambda_X(m, \delta)$ has a Radon-Nikodym derivative with respect to Lebesgue, $f_X(m, \delta) \in \mathbb{C}(\Theta)$, and it is given by

$$f_X(m, \delta) = (I - \varrho a_m)^{-1} [f(\cdot \mid m)].$$

Proof. We only show the proof for part (i) since part (ii) is analogous. Since $\Lambda(m, \delta) \in \mathcal{M}(\Theta)$, its Radon-Nikodym derivative exists; denote it by $f_E(m, \delta)$. Note that, for any $t > \theta_L$,

$$\Lambda(m, \delta)([\theta_L, t]) = \nu([0, t]) + \varrho \int_{\theta_L}^t a_m[f_E(m, \delta)](u) du = \int_{\theta_L}^t \{f_\nu(u) + \varrho a_m[f_E(m, \delta)](u)\} du$$

and, by the fact that a_m maps $L^1(\Theta)$ into $\mathbb{C}(\Theta)$, it follows that $f_E(m, \delta)(\cdot) \equiv f_\nu(\cdot) + \varrho a_m[f_E(m, \delta)](\cdot) \in \mathbb{C}(\Theta)$. Finally, $f_E(m, \delta) = f_\nu + \varrho a_m[f_E(m, \delta)] = (I - \varrho a_m)^{-1}[f_\nu]$. \square

Lemma 9. *Then, for any $m \in \text{int}(\Theta)$ and $\delta \in [0, 1]$,*

$$\frac{d\Lambda(m, \delta)}{dm}(A) = \varrho f_E(m, \delta)(m) \Lambda_X(m, \delta)(A)$$

for any $A \subseteq \Theta$ Borel.

Proof. By Lemma 8, $f_E(m, \delta) = f_\nu + \varrho a_m[f_E(m, \delta)]$ for any $m \in \text{int}(\Theta)$. Let $H : \Theta \times \mathbb{C}(\Theta) \rightarrow \mathbb{C}(\Theta)$ with $H(m, g) \equiv g - f_\nu - \varrho a_m[g]$; observe that $\frac{dH(m, g)}{dm} : \mathbb{R} \rightarrow \mathbb{C}(\Theta)$ with $\frac{dH(m, g)}{dm}[t] = -\varrho f(\cdot | m)g(m)t$ for all $t \in \mathbb{R}$ and $\frac{dH(m, g)}{df} : \mathbb{C}(\Theta) \rightarrow \mathbb{C}(\Theta)$ with $\frac{dH(m, g)}{df}[h] = (I - \varrho a_m)[h]$ for any $h \in \mathbb{C}(\Theta)$. It is easy to see that the mapping $(m, f) \mapsto \|\frac{dH(m, f)}{dm}\|_{L^1} = \varrho f(m)$ is continuous in the sense that, for any f , if f' converges in L^∞ to f , then $\|\frac{dH(m, f')}{dm} - \frac{dH(m, f)}{dm}\|_{L^1} = o(1)$. Also, for any $m_n \rightarrow m$, it follows that

$$\begin{aligned} \frac{\|a_{m_n}[h] - a_m[h]\|_{L^1}}{\|h\|_{L^\infty}} &\leq \left\| \int (1\{\theta \leq m_n\} - 1\{\theta \leq m\}) f(\cdot | \theta) d\theta \right\|_{L^1} \\ &\leq \int |1\{\theta \leq m_n\} - 1\{\theta \leq m\}| d\theta = o(1) \end{aligned}$$

where the first inequality follows from Holder's inequality and the second one from the fact that $\|f(\cdot | \theta)\|_{L^1} = 1$. Moreover, $(\frac{dH(m, g)}{dm})^{-1}$ exists (as a linear bounded operator). So by the Implicit Function Theorem (e.g. Theorem 2.3 in [Ambrosetti and Prodi \(1995\)](#)),

$$\frac{df_E(m, \delta)}{dm} = -(dH(m, f)/df)^{-1} [dH(m, f)/dm] = \varrho (I - \varrho a_m)^{-1} [f(\cdot | m) f_E(m, \delta)(m)]$$

because $\frac{d\phi_m[f_E(m, \delta)](u)}{dm} = \varrho f(u | m) f_E(m, \delta)(m)$. Thus, for any $u \in \Theta$, $\frac{df_E(m, \delta)(u)}{dm} = \varrho (I - \varrho a_m)^{-1} [f(\cdot | m)](u) f_E(m, \delta)(m)$. Integrating at both sides over S , it follows that

$$\begin{aligned} \int_S \frac{df_E(m, \delta)(u)}{dm} du &= \frac{d\Lambda(m, \delta)}{dm}(S) = \varrho f_E(m, \delta)(m) \int_S (I - \varrho a_m)^{-1} [f(\cdot | m)](u) du \\ &= \varrho f_E(m, \delta)(m) \int_S \sum_{j=0}^{\infty} \varrho^j a_m^j [f(\cdot | m)](u) du. \end{aligned}$$

The first equality follows since $\sum_{j=0}^{\infty} \varrho^j a_m^j [f(\cdot | m)](u) \leq K$ for some $K < \infty$, and thus is valid to interchange the integral and the derivative.

Note that for any $A \subseteq \Theta$ Borel,

$$\begin{aligned} \int_A a_m^2[f(\cdot | m)](u) du &= \int_A \int 1\{v \leq m\} a_m[f(\cdot | m)](v) f(u | v) dv du \\ &= \int 1\{v \leq m\} a_m[f(\cdot | m)](v) F(A | v) dv = \Phi_m[\Phi_m[F(\cdot | m)]](A), \end{aligned}$$

and by simple iteration, it follows that

$$\begin{aligned}\frac{d\Lambda(m, \delta)}{dm}(S) &= \varrho f_E(m, \delta)(m) \sum_{j=0}^{\infty} \varrho^j \Phi_m^j [F(\cdot \mid m)](A) \\ &= \varrho f_E(m, \delta)(m) \Lambda_X(m, \delta)(A).\end{aligned}$$

□

Lemma 10. *For any $w > 0$, there exists a unique solution (p, m) to conditions (ii') and (iii'). Moreover, $\bar{q}(p, w, \Lambda(m, \delta)) > 0$; in particular $p > 0$.*

Proof. Throughout the proof, fix $w > 0$. Observe that by Assumption 2(ii), $\nu(\{C(0, w, \theta) > 0\}) > 0$. Also, $\text{supp}(\Lambda(m, \delta)) \supseteq \text{supp}(\nu)$, so $\int C(0, w, \theta) \Lambda(m, \delta)(d\theta) > 0$. This implies that if $\bar{q}(p, w, \Lambda(m, \delta)) = 0$, then $\bar{\pi}(p, w, \Lambda(m, \delta)) < 0 \leq \kappa$, so a (p, m) such that $\bar{q}(p, w, \Lambda(m, \delta)) = 0$ can never be a solution to $\bar{\pi}(p, w, \Lambda(m, \delta)) = \kappa$ (if it exists). Therefore, if the solution exists it be such that $\bar{q}(p, w, \Lambda(m, \delta)) > 0$, in particular, this implies that $p = 0$ cannot be part of a solution.

Henceforth, we thus focus on (p, m) such that $\Lambda(m, \delta)(\{\theta: q(p, w, \theta) > 0\}) > 0$, in particular, we only consider $m \in M \equiv \{m \in \Theta: \exists p: \Lambda(m, \delta)(\{\theta: q(p, w, \theta) > 0\}) > 0\}$. Define the following mappings $m \mapsto p_E(m) \equiv \{p: \bar{\pi}(p, w, \Lambda(m, \delta)) = \kappa\}$ and $p \mapsto m_X(p) \equiv \{m: \bar{\pi}(p, w, \Lambda_X(m, \delta)) = 0\}$ (the dependence on w is left implicit). For the last mapping, it is implicit that if $\bar{\pi}(p, w, \Lambda_X(m, \delta)) < 0$ then $m_X(p) = \theta_L$ and if $\bar{\pi}(p, w, \Lambda_X(m, \delta)) > 0$ then $m_X(p) = \theta_H$. We now characterize these mappings.

The mapping p_E : By Lemma 6(iv), $p \mapsto \pi(p, w, \theta)$ is nondecreasing and increasing over p such that $q(p, w, \theta) > 0$. Therefore, for any $m \in M$, $p \mapsto \bar{\pi}(p, w, \Lambda(m, \delta))$ is increasing. Hence, $p_E(m)$ has at most one element. Moreover, $\bar{\pi}(0, w, \Lambda(m, \delta)) \leq 0$. By Lemma 6(iv) $\liminf_{p \rightarrow \infty} \bar{\pi}(p, w, \Lambda(m, \delta)) = \infty$. Thus, there exists a $\bar{p}(w, m)$ such that $\bar{\pi}(\bar{p}(w, m), w, \Lambda(m, \delta)) > \kappa$. Hence, continuity of $p \mapsto \bar{\pi}(p, w, \Lambda(m, \delta))$ (Lemma 6(iv)) ensures that a solution exists. Therefore $p_E(m)$ is nonempty (and consists of exactly one element, which we still denote $p_E(m)$).

Let $H: \mathbb{R}^2 \rightarrow \mathbb{R}$ where $H(p, m) \equiv \bar{\pi}(p, w, \Lambda(m, \delta))$. By definition, $H(p_E(m), m) = \kappa$ and $\frac{dH(p_E(m), m)}{dp} = \frac{d\bar{\pi}(p_E(m), w, \Lambda(m, \delta))}{dp} > 0$ and $\frac{dH(p_E(m), m)}{dm} = \frac{d\bar{\pi}(p_E(m), w, \Lambda(m, \delta))}{dm}$ which by Lemma 9 exists and is proportional to $\bar{\pi}(p_E(m), w, \Lambda_X(m, \delta))$ (in particular it has the

same sign). By the implicit function theorem,

$$\frac{dp_E(m)}{dm} = - \frac{\frac{d\bar{\pi}(p_E(m), w, \Lambda(m, \delta))}{dm}}{\frac{d\bar{\pi}(p_E(m), w, \Lambda(m, \delta))}{dp}}. \quad (14)$$

The mapping m_X : By definition, $m_X(p)$ is nonempty for any p . By Lemma 7, $\bar{\pi}(p, w, \Lambda_X(m_X(p), \delta)) = V(p, w, m_X(p)) = 0$. By Lemma Lemma 6(v) $m \mapsto V(p, w, m)$ is decreasing, $p \mapsto V(p, w, m)$ is nondecreasing and $(m, p) \mapsto V(p, w, m)$ continuous. Therefore, it is easy to see that $p \mapsto m_X(p)$ is single-valued, continuous and nondecreasing.

For any $m \in \Theta$, let $p_X(m) = \{p: m_X(p) = m\}$. Observe that $p_X(m)$ may not be a singleton because $m_X(\cdot)$ could be “flat” in some parts. However, for any $m \in \Theta$, $p_X(m)$ is closed and convex. Closedness follows from continuity of $p \mapsto V(p, w, m)$. Convexity follows because if $p_1, p_2 \in p_X(m)$ (wlog $p_1 < p_2$) then $V(p_1, w, m) = V(p_2, w, m) = 0$ and since $p \mapsto V(p, w, m)$ is nondecreasing (Lemma Lemma 6(v)) it follows that $0 = V(p, w, m)$ for any $p \in [p_1, p_2]$, so $p \in p_X(m)$. It is easy to see that $p_X(\cdot)$ is increasing in the sense that for any $m_0 < m_1$ and any $p_0 \in p_X(m_0)$, $p_1 \in p_X(m_1)$, $p_0 < p_1$, and it is continuous in the sense that for any $(m_n)_n$ and $(p_n)_n$ such that $m_n \rightarrow m$ and $p_n \in p_X(m_n)$ with $p_n \rightarrow p$ then $p \in p_X(m)$.

Existence of $(p(w), m(w))$: We note that $m(w) = \theta_L$ and $p(w) = p_E(\theta_L)$ if $p_E(m) > p_X(m)$ for any $m \in \Theta$; (ii) $m(w) = \theta_H$ and $p(w) = p_E(\theta_H)$ if $p_E(m) < p_X(m)$ for any $m \in \Theta$. Now consider the case where there exists a m_0 such that $p_E(m) < (>) p_X(m)$ for all $m < (>) m_0$. From the results above $p_E(m_0) \in p_X(m_0)$. Otherwise, either $p_X(m_0)$ is open or nonconvex, which cannot happen.

Uniqueness of $(p(w), m(w))$: To do this, we first establish that for any (p, m) such that $p = p_E(m)$ and m is such that $\bar{\pi}(p_E(m), w, \Lambda_X(m, \delta)) = 0$, then $\frac{d^2 p_E(m)}{dm^2} \geq 0$. To do this, we employ display 14, to show that

$$\frac{d^2 p_E(m)}{dm^2} = - \frac{\frac{d^2 \bar{\pi}(p_E(m), w, \Lambda(m, \delta))}{dm^2}}{\frac{d\bar{\pi}(p_E(m), w, \Lambda(m, \delta))}{dp}} + \frac{\frac{d\bar{\pi}(p_E(m), w, \Lambda(m, \delta))}{dm} \frac{d^2 \bar{\pi}(p_E(m), w, \Lambda(m, \delta))}{dp dm}}{\left(\frac{d\bar{\pi}(p_E(m), w, \Lambda(m, \delta))}{dp} \right)^2}.$$

Since by Lemma 9, $\frac{d\bar{\pi}(p_E(m), w, \Lambda(m, \delta))}{dm}$ has the same sign as $\bar{\pi}(p_E(m), w, \Lambda_X(m, \delta))$ and m is such that $\bar{\pi}(p_E(m), w, \Lambda_X(m, \delta)) = 0$, it follows that the second term in the RHS

is zero. Therefore, it is sufficient to show that $\frac{d^2 \bar{\pi}(p_E(m), w, \Lambda(m, \delta))}{dm^2} \leq 0$. By Lemma 9,

$$\begin{aligned} \frac{d^2 \bar{\pi}(p_E(m), w, \Lambda(m, \delta))}{dm^2} &= \frac{d [\varrho f_E(m, \delta)(m) \bar{\pi}(p_E(m), w, \Lambda_X(m, \delta))]}{dm} \\ &= \varrho \frac{d [f_E(m, \delta)(m)]}{dm} \bar{\pi}(p_E(m), w, \Lambda_X(m, \delta)) \\ &\quad + \varrho f_E(m, \delta)(m) \frac{d [\bar{\pi}(p_E(m), w, \Lambda_X(m, \delta))]}{dm}. \end{aligned}$$

The first term in the RHS is zero because $\bar{\pi}(p_E(m), w, \Lambda_X(m, \delta)) = 0$. So it boils down to show that $\frac{d [\bar{\pi}(p_E(m), w, \Lambda_X(m, \delta))]}{dm} \leq 0$. By Lemma 7 and the fact that $dp_E(m)/dm = 0$, it suffices to show that $\frac{dV_m(p, w, m)}{dm} \leq 0$ at (p, m) such that $p = p_E(m)$ and m is such that $\bar{\pi}(p_E(m), w, \Lambda_X(m, \delta)) = 0$. Since $V_m(p, w, m) = \pi(p, w, m) + \varrho M[V_m(p, w, \cdot)](m)$, so

$$\frac{dV_m(p, w, m)}{dm} = \frac{d\pi(p, w, m)}{dm} + \frac{d \int V_m(p, w, \theta') f(\theta' | t) d\theta'}{dt} \Big|_{t=m} + \varrho \frac{\partial M[V_m(p, w, \cdot)](m)}{\partial m}$$

where $\frac{\partial M[V_m(p, w, \cdot)](m)}{\partial m}$ is the derivative of the function $m \mapsto \frac{\partial M[V_m(p, w, \cdot)](\theta)}{\partial m}$ with respect to m , evaluated at $\theta = m$. Observe that, for any $\theta \in \Theta$,

$$\begin{aligned} M[V_m(p, w, \cdot)](\theta) &= M[\pi(p, w, \cdot)](\theta) + \varrho \int_{\theta_L}^m M[V_m(p, w, \cdot)](\theta') F(d\theta' | \theta) \\ &= M[\pi(p, w, \cdot)](\theta) + \varrho (\Phi_m \circ M)[V_m(p, w, \cdot)](\theta). \end{aligned}$$

Thus, by Leibniz's rule, for any $\theta \in \Theta$,

$$\frac{\partial M[V_m(p, w, \cdot)](\theta)}{\partial m} = \varrho M[V_m(p, w, \cdot)](m) f(m | \theta) + \varrho \Phi_m \left[\frac{\partial M[V_m(p, w, \cdot)]}{\partial m} \right](\theta)$$

and thus $\frac{\partial M[V_m(p, w, \cdot)](\theta)}{\partial m} = \varrho M[V_m(p, w, \cdot)](m) \times (I - \varrho \Phi_m)^{-1} [f(m | \cdot)](\theta)$. By definition of (p, m) and Lemma 3, $M[V_m(p, w, \cdot)](m) = \bar{\pi}(p, w, \Lambda_X(m, \delta)) = M[V(p, w, \cdot)](m)$, then $\frac{\partial M[V_m(p, w, \cdot)](\theta)}{\partial m} = 0$. Therefore, under Assumption 3(i)

$$\frac{dV_m(p, w, m)}{dm} = \frac{d\pi(p, w, m)}{dm} + \int V_m(p, w, \theta') \frac{df(\theta' | m)}{dm} d\theta'.$$

By Lemma 6(iv), $\frac{d\pi(p, w, m)}{dm} \leq 0$. By Lemma 6(v) $\theta' \mapsto V_m(p, w, \theta')$ is decreasing so under assumption 3(ii) $t \mapsto \int V_m(p, w, \theta') f(\theta' | t) d\theta'$ is decreasing; hence $\frac{d \int V_m(p, w, \theta') f(\theta' | t) d\theta'}{dt} \leq 0$ at $t = m$. Thus $\frac{dV_m(p, w, m)}{dm} \leq 0$, and this shows that $\frac{d^2 p_E(m)}{dm^2} \geq 0$.

Since p_E is differentiable, the fact that for any m such that $\bar{\pi}(p_E(m), w, \Lambda_X(m, \delta)) = 0$ (which is equivalent to $\frac{dp_E(m)}{dm} = 0$), then $\frac{d^2 p_E(m)}{dm^2} \geq 0$, implies that there exists at most one $m \in \text{Int}(\Theta)$ such that $\frac{dp_E(m)}{dm} = 0$ and moreover, such m is a local minimizer of p_E . Therefore, there exists at most one pair $(p(w), m(w))$ with $m \in \text{Int}(\Theta)$ such that $\bar{\pi}(p(w), w, \Lambda_X(m(w), \delta)) = 0$ and $\bar{\pi}(p(w), w, \Lambda(m(w), \delta)) = \kappa$. Therefore, $(p(w), m(w))$ is unique and characterized by: (i) $m(w) = \theta_L$ and $p(w) = p_E(\theta_L)$ if $p_E(m) > p_X(m)$ for any $m \in \Theta$; (ii) $m(w) = \theta_H$ and $p(w) = p_E(\theta_H)$ if $p_E(m) < p_X(m)$ for any $m \in \Theta$; and (iii) $(m(w), p(w))$ such that $\bar{\pi}(p(w), w, \Lambda_X(m(w), \delta)) = 0$ and $\bar{\pi}(p(w), w, \Lambda(m(w), \delta)) = \kappa$, otherwise.

[Proof of Lemma 4] Throughout the proof, fix an arbitrary $w > 0$. Let $\mathcal{S}(w) \equiv \{(p, m) \in \mathbb{R}_{++} \times \Theta : \text{satisfy conditions (ii')} \text{ and (iii')}\}$ (by Lemma 10 it is nonempty and a singleton, also the solution satisfies $\bar{q}(p, w, \Lambda(m, \delta)) > 0$). For any $m \in \Theta$, let $p^*(m, w) = \arg \min_{p \geq 0} \bar{A}C(p, w, \Lambda(m, \delta))$ and let $m^*(w) = \arg \min_{m \in \Theta} \bar{A}C(p^*(m), w, \Lambda(m, \delta))$. The result is proven if we show that $\mathcal{S} \subseteq \{p^*(m^*(w), w), m^*(w)\}$.

In order to do this, let $\mathcal{F}(m, w) \equiv \left\{ p \geq 0 : \frac{d\bar{A}C(p, w, \Lambda(m, \delta))}{dp} = 0 \text{ and } p = \bar{A}C(p, w, \Lambda(m, \delta)) \right\}$ and $\mathcal{S}(m, w) \equiv \{p \geq 0 : \bar{\pi}(p(m, w), w, \Lambda(m, \delta)) = \kappa\}$. Suppose that (we show these below): (a) $\mathcal{S}(m, w) \subseteq \mathcal{F}(m, w)$, (b) $\mathcal{F}(m, w)$ contains at most one element, (c) $\mathcal{F}(m, w) = p^*(m, w)$ and (d) $m \mapsto p^*(m, w)$ is continuous differentiable.

Let $(p, m) \in \mathcal{S}$, then $p \in \mathcal{S}(m, w)$ and by (a)-(c) it holds that $p \in p^*(m, w)$. So the desired result follows if $m \in m^*(w)$. To show this last relationship we study the mapping $w \mapsto m^*(w)$. First, by compactness of Θ and continuity of $m \mapsto \bar{A}C(p^*(m, w), w, \Lambda(m, \delta))$ (continuity of $m \mapsto p^*(m, w)$ follows from (d)) $m^*(w)$ is nonempty. Moreover, by the envelope theorem and (c)-(d), the minimizer must satisfy the following

$$\frac{d\bar{A}C(p^*(m^*(w), w), w, \Lambda(m^*(w), \delta))}{dm} \begin{cases} = 0 & \text{if } m^*(w) \in (\theta_L, \theta_H) \\ \geq 0 & \text{if } m^*(w) = \theta_H \\ \leq 0 & \text{if } m^*(w) = \theta_L \end{cases} . \quad (15)$$

We note that $\frac{d\Lambda(m, \delta)}{dm}$ is proportional to $\Lambda_X(m, \delta)$ and that both $\theta \mapsto q(p, w, \theta)$ and $q \mapsto C(q, w, \theta)$ are continuous (see Lemma 6(i)-(ii)). Thus, by the DCT we can interchange integration and differentiation and obtain that, for any p such that $\bar{q}(p, w, \Lambda(m, \delta)) >$

0,

$$\begin{aligned} \frac{d\bar{A}C(p, w, \Lambda(m, \delta))}{dm} &= \frac{\int_{\Theta} C(q(p, w, \theta), w, \theta) \frac{d\Lambda(m, \delta)}{dm}(d\theta)}{\bar{q}(p, w, \Lambda(m, \delta))} \\ &\quad - \frac{\bar{C}(p, w, \Lambda(m, \delta)) \int_{\Theta} q(p, w, \theta) \frac{d\Lambda(m, \delta)}{dm}(d\theta)}{\bar{q}(p, w, \Lambda(m, \delta))^2}. \end{aligned}$$

Note that $\bar{\pi}(p^*(m, w), w, \Lambda(m, \delta)) = \kappa \iff \bar{A}C(p^*(m, w), w, \Lambda(m, \delta)) = p^*(m, w)$ for any m , so the display above, for $p = p^*(m, w)$, implies

$$\begin{aligned} \frac{d\bar{A}C(p^*(m, w), w, \Lambda(m, \delta))}{dm} &= \frac{\int_{\Theta} C(q(p^*(m, w), w, \theta), w, \theta) \frac{d\Lambda(m, \delta)}{dm}(d\theta)}{\bar{q}(p^*(m, w), w, \Lambda(m, \delta))} \\ &\quad - p^*(m, w) \frac{\int_{\Theta} q(p^*(m, w), w, \theta) \frac{d\Lambda(m, \delta)}{dm}(d\theta)}{\bar{q}(p^*(m, w), w, \Lambda(m, \delta))}. \end{aligned}$$

By Lemma 9,

$$\int_{\Theta} C(q(p, w, \theta), w, \theta) \frac{d\Lambda(m, \delta)}{dm}(d\theta) = \varrho f_E(m, \delta)(m) \int_{\Theta} C(q(p, w, \theta), w, \theta) \Lambda_X(m, \delta)(d\theta)$$

and similarly for $\int_{\Theta} q(p, w, \theta) \frac{d\Lambda(m, \delta)}{dm}(d\theta)$. Therefore,

$$\begin{aligned} \frac{d\bar{A}C(p^*(m, w), w, \Lambda(m, \delta))}{dm} &= \frac{\varrho f_E(m, \delta)(m)}{\bar{q}(p^*(m, w), w, \Lambda(m, \delta))} \\ &\quad \times \{ \bar{C}(p^*(m, w), w, \Lambda_X(m, \delta)) - p^*(m, w) \bar{q}(p^*(m, w), w, \Lambda_X(m, \delta)) \} \\ &= \frac{\varrho f_E(m, \delta)(m)}{\bar{q}(p^*(m, w), w, \Lambda(m, \delta))} \bar{\pi}(p^*(m, w), w, \Lambda_X(m, \delta)), \end{aligned}$$

where the second line follows from the definition $\bar{\pi}(p^*(m, w), w, \Lambda_X(m, \delta))$. Observe that $f_E(m, \delta) > 0$ because $f_{\nu} > 0$ by Assumption 5(i) and $\text{supp}(\Lambda(m, \delta)) \supseteq \text{supp}(\nu)$. Therefore $\text{sign} \left\{ \frac{d\bar{A}C(p^*(m, w), w, \Lambda(m, \delta))}{dm} \right\} = \text{sign} \{ \bar{\pi}(p^*(m, w), w, \Lambda_X(m, \delta)) \}$. Therefore, display 15 is equivalent to

$$\begin{aligned} \bar{\pi}(p^*(m^*(w), w), w, \Lambda_X(m^*(w), \delta)) &= 0 & \text{if } m^*(w) \in (\theta_L, \theta_H) \\ \bar{\pi}(p^*(m^*(w), w), w, \Lambda_X(m^*(w), \delta)) &\leq 0 & \text{if } m^*(w) = \theta_H \\ \bar{\pi}(p^*(m^*(w), w), w, \Lambda_X(m^*(w), \delta)) &\geq 0 & \text{if } m^*(w) = \theta_L \end{aligned} \quad (16)$$

Therefore, $m \in m^*(w)$ as desired. Finally, we note that by (a)-(c) $p = \bar{A}C(p, w, \Lambda(m, \delta)) >$

0.

We now show (a)-(d). To show (a), take $p \in \mathcal{S}(m, w)$ (which exists by 10). For such p , $\bar{q}(p, w, \Lambda(m, \delta)) > 0$. Since $\bar{AC}(\cdot, w, \Lambda(m, \delta))$ is continuously differentiable, the FONC $\frac{d\bar{AC}(p, w, \Lambda(m, \delta))}{dp} = 0$ is given by

$$\frac{\int_{\Theta} \frac{dC(q(p, w, \theta), w, \theta)}{dq} \frac{dq(p, w, \theta)}{dp} \Lambda(m, \delta)(d\theta)}{\bar{q}(p, w, \Lambda(m, \delta))} - \frac{\bar{C}(p, w, \Lambda(m, \delta)) \int_{\Theta} \frac{dq(p, w, \theta)}{dp} \Lambda(m, \delta)(d\theta)}{\bar{q}(p, w, \Lambda(m, \delta))^2} = 0.$$

By the proof of Lemma 6(ii), $q(p, w, \theta) \left(\frac{dC(q(p, w, \theta), w, \theta)}{dq} - p \right) = 0$ for any $p > 0$, so by total differentiation it follows that $\frac{dq(p, w, \theta)}{dp} \left(\frac{dC(q(p, w, \theta), w, \theta)}{dq} - p \right) + q(p, w, \theta) \left(\frac{d^2 C(q(p, w, \theta), w, \theta)}{dq dp} - 1 \right) = 0$. Thus, if $q(p, w, \theta) = 0$, then $\frac{dq(p, w, \theta)}{dp} \left(\frac{dC(q(p, w, \theta), w, \theta)}{dq} - p \right) = 0$; so $q(p, w, \theta) \left(\frac{dC(q(p, w, \theta), w, \theta)}{dq} - p \right) = 0$ implies $\frac{dq(p, w, \theta)}{dp} \frac{dC(q(p, w, \theta), w, \theta)}{dq} = p \frac{dq(p, w, \theta)}{dp}$. By applying this fact and the fact that we can interchange integration and differentiation,¹⁸ the FONC becomes

$$0 = p \frac{\int_{\Theta} \frac{dq(p, w, \theta)}{dp} \Lambda(m, \delta)(d\theta)}{\bar{q}(p, w, \Lambda(m, \delta))} - \frac{\bar{C}(p, w, \Lambda(m, \delta)) \int_{\Theta} \frac{dq(p, w, \theta)}{dp} \Lambda(m, \delta)(d\theta)}{\bar{q}(p, w, \Lambda(m, \delta))^2}.$$

Therefore, the FOC of \bar{AC} with respect to p implies

$$(\bar{q}(p, w, \Lambda(m, \delta))p - \bar{C}(p, w, \Lambda(m, \delta))) \int_{\Theta} \frac{dq(p, w, \theta)}{dp} \Lambda(m, \delta)(d\theta) = 0.$$

It is easy to see that this display implies that $\{\bar{\pi}(p, w, \Lambda(m, \delta)) - \kappa\} \int_{\Theta} \frac{dq(p, w, \theta)}{dp} \Lambda(m, \delta)(d\theta) = 0$. Since $\bar{q}(p, w, \Lambda(m, \delta)) > 0$ it follows by Lemma 6(ii) that $\int_{\Theta} \frac{dq(p, w, \theta)}{dp} \Lambda(m, \delta)(d\theta) > 0$, so $\bar{\pi}(p, w, \Lambda(m, \delta)) = \kappa$. Therefore, for any $m \in \Theta$, $\frac{d\bar{AC}(p^*(m, w), w, \Lambda(m, \delta))}{dp} = 0$ is equivalent to

$$\bar{\pi}(p^*(m, w), w, \Lambda(m, \delta)) = \kappa.$$

It is straightforward to show that is also equivalent to $p = \bar{AC}(p, w, \Lambda(m, \delta))$. Therefore, $p \in \mathcal{F}(m, w)$ as desired.

We now show (b) by contradiction. That is, suppose there exist a p_1 $p \neq p_1 \in \mathcal{F}(m, w)$. We note that $\lim_{p \rightarrow 0} \bar{AC}(p, w, \Lambda(m, \delta)) = \infty$, so this and the definition of $\mathcal{F}(m, w)$ imply that such p_1 is such that: either $\frac{d\bar{AC}(p_1, w, \Lambda(m, \delta))}{dp} = 0$ and $p_1 >$

¹⁸This fact follows from the DCT and the fact that $\frac{dq(p, w, \theta)}{dp} \leq 1 / \frac{d^2 C(q(p, w, \theta), w, \theta)}{dq^2}$ (see the proof of Lemma 6(ii)) and that $\theta \mapsto \frac{d^2 C(q(p, w, \theta), w, \theta)}{dq^2}$ is continuous and strictly positive (see the proof of Lemma 6(i)).

$\bar{AC}(p_1, w, \Lambda(m, \delta))$, or $\frac{d\bar{AC}(p_1, w, \Lambda(m, \delta))}{dp} > 0$ and $p_1 = \bar{AC}(p_1, w, \Lambda(m, \delta))$; both being contradictions. Therefore, (b) holds.

Regarding (c), we first note that $\liminf_{p \rightarrow \infty} \bar{AC}(p, w, \Lambda(m, \delta)) = \infty$. To show this, note that by convexity of $C(\cdot, w, \theta)$ (Lemma 6(i)) and the fact that

$$\begin{aligned} \bar{AC}(p, w, \Lambda(m, \delta)) &\geq \frac{\int C(0, w, \theta) \Lambda(m, \delta)(d\theta)}{\bar{q}(p, \Lambda(m, \delta))} + \frac{\int \frac{dC(q(p, w, \theta), w, \theta)}{dq} q(p, w, \theta) \Lambda(m, \delta)(d\theta)}{\bar{q}(p, w, \Lambda(m, \delta))} \\ &= \frac{\int C(0, w, \theta) \Lambda(m, \delta)(d\theta)}{\bar{q}(p, w, \Lambda(m, \delta))} + p \end{aligned}$$

where the second line follows from the fact that $\frac{dC(q(p, w, \theta), w, \theta)}{dq} q(p, w, \theta) = pq(p, w, \theta)$ (see 6(ii)). Since $\frac{\int C(0, w, \theta) \Lambda(m, \delta)(d\theta)}{\bar{q}(p, w, \Lambda(m, \delta))} \geq 0$, the previous display implies that $\liminf_{p \rightarrow \infty} \bar{AC}(p, w, \Lambda(m, \delta)) = \infty$. We already establish that $\lim_{p \rightarrow 0} \bar{AC}(p, w, \Lambda(m, \delta)) = \infty$ and that $\bar{AC}(\cdot, w, \Lambda(m, \delta))$ is continuously differentiable. So $p^*(m, w)$ is nonempty and must be such that $\mathcal{F}(m, w) = p^*(m, w)$.

The claim (d) follows by the implicit function theorem, and for any $m \in \text{int}(\Theta)$

$$\frac{dp^*(m, w)}{dm} = \frac{d\bar{AC}(p^*(m, w), w, \Lambda(m, \delta))}{dm}.$$

□

A.5 Proof of Lemma 5

Proof. [Proof of Lemma 5] The case where W is constant is trivial, so we consider the case where W is nondecreasing (with parts being increasing). For this case, we impose $\pi(p, w, \theta) = h(\theta)g(p, w) - FC(\theta)$ with g differentiable (Assumption 6). By construction of π , it follows that $h(\theta)g(p, w) = pf(l(p, w, \theta), \theta) - wl(p, w, \theta)$ for any (p, w, θ) . Taking derivatives with respect to w and using the fact that $p \frac{df(l(p, w, \theta), \theta)}{dl} - w = 0$, it follows that

$$-h(\theta) \frac{dg(p, w)}{dw} = l(p, w, \theta).$$

Similarly, taking derivatives with respect to p ,

$$h(\theta) \frac{dg(p, w)}{dp} = f(l(p, w, \theta), \theta) = q(p, w, \theta).$$

Also, note that

$$\frac{\frac{dg(p,w)}{dp}}{\frac{dg(p,w)}{dw}} = -\frac{f(l(p,w,\theta),\theta)}{l(p,w,\theta)}$$

which means that $\frac{f(l(p,w,\theta),\theta)}{l(p,w,\theta)}$ does not depend on $\theta \in \Theta$; henceforth $\varphi(p,w) \equiv \frac{f(l(p,w,\theta),\theta)}{l(p,w,\theta)}$; note that φ is homogenous of degree 0 because $l(\cdot, \cdot, \theta)$ is (see Lemma 6(iii)). We also note that, under assumption 2 and the fact that $w/p \mapsto l(1, w/p, \theta)$ is nonincreasing and decreasing when $l(p,w,\theta) > 0$ (see Lemma 6(iii)), it follows that $\frac{w}{p} \mapsto \varphi(p,w)$ is nondecreasing and increasing when $l(p,w,\theta) > 0$. Hence

$$W\left(Q\frac{\bar{l}(p,w,\Lambda(m,\delta))}{\bar{q}(p,w,\Lambda(m,\delta))}\right) = W(-Q\varphi(1,w/p))$$

and it is nonincreasing as a function of w/p . Suppose $w \mapsto w/p(w)$ is nondecreasing (it is shown below), thus $w \mapsto W(-Q\varphi(1,w/p(w)))$ is nonincreasing. Hence, there exists a unique solution to $w = W(-Q\varphi(1,w/p(w)))$ which we denote by $\hat{w}(Q)$. Since W is continuous (by Assumption 1(ii)) and so is $w \mapsto p(w)$ (see the proof of Lemma 4), it follows that $Q \mapsto \hat{w}(Q)$ is continuous and nondecreasing.

We now show that $w \mapsto w/p(w)$ is nondecreasing. It follows that, for any $p > 0$, w and θ , $\pi(p,w,\theta) = p\pi(1,w/p,\theta)$ and it readily follows that V inherits this property, i.e., $V(p,w,\theta) = pV(1,w/p,\theta)$. Also, by Lemma 6(v), $\frac{w}{p} \mapsto V(1,w/p,\theta)$ is nonincreasing. These results and condition (ii) in Definition 2 imply that $\int p(w)V(1,w/p(w),\theta)\nu(d\theta) = \kappa$. So, if $w \mapsto p(w)$ is nondecreasing, the previous equation and the fact that $\frac{w}{p} \mapsto V(1,w/p,\theta)$ is nonincreasing imply that $w \mapsto w/p(w)$ is nondecreasing. The fact that $w \mapsto p(w)$ is nondecreasing readily follows from the equation $\int V(p(w),w,\theta)\nu(d\theta) = \kappa$ and analogous arguments. \square

A.6 Proof of Proposition 2

Proof. One option is to set $Q = n = 0$ and obtain zero surplus. We know this option is not the best, since the planner can always replicate the equilibrium which, by Corollary 1, is characterized by positive aggregate production and, therefore, positive surplus. We now consider if the planner can do better by choosing some other positive production level.

Step 1. Choose production allocation. Fix $Q > 0$, $m \in \Theta$, $n > 0$. Find optimal

allocation to minimize cost. It is easy to see that we want to equalize MC among all firms that produce positive quantity. So this is equivalent to choosing a marginal cost, which we denote by $p(Q, m, n)$ with the property that

$$Q = \bar{q}(p(Q, m, n), \Lambda(m, 1))n. \quad (17)$$

Step 2. Choose $n > 0$. Fix $Q > 0$ and $m \in \Theta$ and choose n to minimize $n(\int_{\Theta} C(q(p(Q, m, n), \theta), \theta) \Lambda(m, 1)(d\theta) + \kappa)$ subject to constraint (17). It is easy to see, using the fact that $\frac{dC}{dq}(q(p(Q, m, n), \theta), \theta) = p(Q, m, n)$ for all θ such that $q(p(Q, m, n), \theta) > 0$, that the solution n^* is such that

$$p(Q, m, n^*) = \bar{AC}(p(Q, m, n^*), \Lambda(m, 1)) = p^*(m),$$

where $p^*(m) \equiv \arg \min_p \bar{AC}(p, \Lambda(m, 1))$. The minimized total cost is then $Q \cdot \bar{AC}(p^*(m), \Lambda(m, 1))$.

Step 3. Choose $m \in \Theta$. Fix $Q > 0$ and choose $m \in \Theta$ to minimize the total cost minimized in step 2, $Q \cdot \bar{AC}(p^*(m), \Lambda(m, 1))$. Denote the solution by $m^* = \arg \min_m \bar{AC}(p^*(m), \Lambda(m, 1))$.

Step 4. The final step is to find the aggregate amount of production Q that maximizes total surplus,

$$\max_Q \int_0^Q P^d(\tilde{Q}) d\tilde{Q} - Q \cdot \bar{AC}(p^*(m^*), \Lambda(m^*, 1)).$$

The optimal aggregate quantity Q^* is the unique solution to

$$P^d(Q^*) = \bar{AC}(p^*(m^*), \Lambda(m^*, 1)) = p^*. \quad (18)$$

By equation (18), $Q^d(p^*) = Q^*$, where, by equation (17), $Q^* = \bar{q}(p^*, \Lambda(m^*, 1))n^* = Q^s(p^*; n^*, m^*)$; therefore, the allocation that maximizes steady-state surplus is given by the allocation induced by $\langle p^*, n^*, m^* \rangle$, where $(p^*, m^*) = \arg \min_{p, m} \bar{AC}(p, \Lambda(m, 1))$. \square

A.7 Proof of Proposition 3

Proof. For each $\delta \in [0, 1)$, let $\langle p_{\delta}^e, n_{\delta}^e, m_{\delta}^e \rangle$ denote the corresponding LRCE. We first show that for any $\delta_1 < \delta_2$, then $p_{\delta_2}^e < p_{\delta_1}^e$. It follows that for each $\delta \in [0, 1)$, $\int V_{\delta}(p_{\delta}^e, \theta) \nu(d\theta) = \kappa$ (the notation V_{δ} makes the dependence of V on δ explicit). By

Lemma 6(vi), $\delta \mapsto V_\delta(p, w, \theta)$ is nondecreasing. Moreover, by inspection of the proof, it also follows that $\delta \mapsto V_\delta(p, w, \theta)$ is increasing for all $\theta < m_\delta^e$. Since ν has full support (Assumption 5), this implies that $\delta \mapsto \int V_\delta(p, \theta) \nu(d\theta)$ is increasing. Since p_δ^e solves $\int V_\delta(p_\delta^e, \theta) \nu(d\theta) = \kappa$, the former fact and the fact that $p \mapsto V(p, \theta)$ is nondecreasing by Lemma 6(v) imply that $\delta \mapsto p_\delta^e$ is decreasing.

The result that $\delta \mapsto p_\delta^e$ is decreasing readily implies that $p^* < p_\delta^e$ for any $\delta \in [0, 1)$. This fact implies that $Q^d(p^*) > Q^d(p_\delta^e)$ under Assumption 1, and also implies that $q(p^*, \cdot) \leq q(p_\delta^e, \cdot)$ by Lemma 6(ii); in fact the inequality is strict for any θ such that $q(p_\delta^e, \theta) > 0$. \square

A.8 Proof of Proposition 4

Proof. Let $(p^*, m^*) = \arg \min_{p, m} \bar{AC}(p, \Lambda(m, 1))$ and define, for all $\xi > 0$, $N(\xi) \equiv \{p : p \in (p^*, p^* + \xi), \bar{\pi}(p, m^*, \Lambda(m^*, \delta)) < \kappa < \bar{\pi}(p, m^*, \Lambda(m, 1)), \bar{\pi}(p, m^*, \Lambda_X(m^*, \delta)) < 0, Q^d(p) > 0\}$, a set that is nonempty due Lemmas 4 and 6(iv), by Proposition 3, and by the fact that Q^d is continuous and $Q^d(p^*) > 0$. Fix $\varepsilon > 0$. By continuity of $S(\cdot)$, $Q^d(\cdot)$, and $Q^s(\cdot)$, there exists $\xi(\varepsilon) > 0$ such that, for all $p_\varepsilon \in N(\xi(\varepsilon))$ and $n_\varepsilon > 0$ such that $Q^d(p_\varepsilon) = Q^s(p_\varepsilon; n_\varepsilon, m^*) > 0$, then $S(p_\varepsilon, n_\varepsilon, m^*) \geq \mathcal{S}^* - \varepsilon$. We will show that there is a tax policy under which $\langle p_\varepsilon, m^*, n_\varepsilon \rangle$ is the unique LRCE. Let the tax policy $\mathcal{T}_\varepsilon = \langle \tau_\varepsilon, S_\varepsilon^E, S_\varepsilon^X \rangle$ be such that

$$\tau_\varepsilon = \frac{\kappa - (\bar{\pi}(p_\varepsilon, \Lambda(m^*, \delta)) + (\Lambda(m^*, 1)(\Theta) - 1)\bar{\pi}(p_\varepsilon, m^*, \Lambda_X(m^*, 1)))}{\bar{\pi}(p_\varepsilon, \Lambda(m^*, 1)) - (\bar{\pi}(p_\varepsilon, \Lambda(m^*, \delta)) + (\Lambda(m^*, 1)(\Theta) - 1)\bar{\pi}(p_\varepsilon, m^*, \Lambda_X(m^*, 1)))},$$

$S_\varepsilon^E = \kappa - (1 - \tau_\varepsilon)\bar{\pi}(p_\varepsilon, \Lambda(m^*, \delta))$, and $S_\varepsilon^X = -(1 - \tau_\varepsilon)\bar{\pi}(p_\varepsilon, \Lambda_X(m^*, \delta))$. The facts that $p_\varepsilon \in N(\xi(\varepsilon))$ and $\Lambda(m^*, 1)(\Theta) \geq \nu(\Theta) = 1$ imply that $\tau_\varepsilon \in (0, 1)$. By construction, condition (iv) in Definition 3 is satisfied. By the same arguments as in the proof of Lemma 3, it follows that conditions (ii) and (iii) in Definition 3 are equivalent to $(1 - \tau_\varepsilon)\bar{\pi}(p, \Lambda(m, \delta)) + S_\varepsilon^E = \kappa$ and $(1 - \tau_\varepsilon)\bar{\pi}(p, \Lambda_X(m, \delta)) = (\geq)(\leq) - S_\varepsilon^X$ if $m \in (\theta_L, \theta_H)$ ($m = \theta_H$) ($m = \theta_L$). Given the definition of $\tau_\varepsilon, S_\varepsilon^E, S_\varepsilon^X$, it follows that these two conditions are equivalent to $\bar{\pi}(p, \Lambda(m, \delta)) = \bar{\pi}(p_\varepsilon, \Lambda(m^*, \delta))$ and $\bar{\pi}(p, \Lambda_X(m, \delta)) = \bar{\pi}(p_\varepsilon, \Lambda_X(m^*, \delta))$. Thus, we have a system of equations in (p, m) and, by the same arguments provided in Lemma 10, there is a unique solution to these equations. Moreover, it is obvious that this solution is (p_ε, m^*) . Thus, $(p_\varepsilon, m^*, n_\varepsilon)$ is the unique LRCE with tax policy \mathcal{T}_ε . Finally, note that $\lim_{\varepsilon \rightarrow 0} \xi(\varepsilon) = 0$ implies that $\lim_{\varepsilon \rightarrow 0} p_\varepsilon = p^*$, and,

therefore, by continuity of $\bar{\pi}(\cdot, \Lambda(m^*, 1))$, $\lim_{\varepsilon \rightarrow 0} \bar{\pi}(p_\varepsilon, \Lambda(m^*, 1)) = \bar{\pi}(p^*, \Lambda(m^*, 1)) = \kappa$. Therefore, $\lim_{\varepsilon \rightarrow 0} \tau_\varepsilon = 1$ and, by the fact that $\bar{\pi}(\cdot, \Lambda(m^*, \delta))$ and $\bar{\pi}(p_\varepsilon, \Lambda_X(m^*, \delta))$ are bounded, $\lim_{\varepsilon \rightarrow 0} S_\varepsilon^E = \kappa$ and $\lim_{\varepsilon \rightarrow 0} S_\varepsilon^X = 0$. \square

A.9 Proof of Proposition 5

Proof. Suppose not. That is, suppose there exists a subsequence, which we still denote as $(\tau_j)_j$, such that $\tau_\infty \equiv \lim_{j \rightarrow \infty} \tau_j < 1$. We first show that $\lim_{j \rightarrow \infty} (p_j^e, m_j^e) \neq (p^*, m^*)$. To do this, note that by Condition (iv) in Definition 4, (a) $\int \tau_j \pi(p_j^e, \theta) \Lambda(m_j^e, 1) (d\theta) \geq (S_j^E + (\Lambda(m_j, 1)(\Theta) - 1)S_j^X)$ for each j . By following the same steps as in Lemma 7, one can cast conditions (ii) and (iii) in Definition 4 as (b) $(1 - \tau_j)\bar{\pi}(p_j^e, \Lambda(m_j^e, \delta)) + S_j^E = \kappa$ and (c) $\bar{\pi}(p_j^e, \Lambda_X(m_j^e, \delta)) + S_j^X = 0$ if $m^e \in (\theta_L, \theta_H)$, ≥ 0 if $m^e = \theta_H$, and ≤ 0 if $m^e = \theta_L$.

After some straightforward manipulations, one can show that expressions (a)-(c) imply

$$\kappa \leq (1 - \tau_j)\bar{\pi}(p_j^e, \Lambda(m_j^e, \delta)) + (\Lambda(m_j, 1)(\Theta) - 1)\bar{\pi}(p_j^e, \Lambda_X(m_j^e, \delta)) + \tau_j\bar{\pi}(p_j^e, \Lambda(m_j^e, 1)).$$

Suppose that $\lim_{j \rightarrow \infty} (p_j^e, m_j^e) = (p, m)$ (if necessary, going to a further sub-sequence). Then by continuity of $\pi(\cdot, \theta)$, it follows that

$$\kappa \leq (1 - \tau_\infty)\bar{\pi}(p, \Lambda(m, \delta)) + (\Lambda(m, 1)(\Theta) - 1)\bar{\pi}(p, \Lambda_X(m, \delta)) + \tau_\infty\bar{\pi}(p, \Lambda(m, 1)).$$

By Lemma 7, $\bar{\pi}(p, \Lambda(m, \delta)) = \int V_\delta(p, \theta) \nu(d\theta)$ and by Lemma 6(vi), $\delta \mapsto V_\delta(p, \theta)$ is non-decreasing (increasing for all $\theta < m$), thus $(1 - \tau_\infty)\bar{\pi}(p, \Lambda(m, \delta)) + \tau_\infty\bar{\pi}(p, \Lambda(m, 1)) < \bar{\pi}(p, \Lambda(m, 1))$ and $\bar{\pi}(p, \Lambda_X(m, \delta)) \leq \bar{\pi}(p, \Lambda_X(m, 1))$; the strict inequality follows because ν (and thus Λ) has full support. Hence

$$\kappa < (\Lambda(m, 1)(\Theta) - 1)\bar{\pi}(p, \Lambda_X(m, 1)) + \bar{\pi}(p, \Lambda(m, 1)).$$

Note that $(\Lambda(m^*, 1)(\Theta) - 1)\bar{\pi}(p^*, \Lambda_X(m^*, 1)) \geq 0$, because $\bar{\pi}(p^*, \Lambda_X(m^*, 1)) \geq 0$ if $m^* \in (\theta_L, \theta_H]$ and $\Lambda(\theta_L, 1)(\Theta) = \nu(\Theta) = 1$, and $\bar{\pi}(p^*, \Lambda(m^*, 1)) = \kappa$. Therefore, the previous display shows that $(p, m) \neq (p^*, m^*)$. That is, under $\lim_{j \rightarrow \infty} \tau_j < 1$, it follows that $\lim_{j \rightarrow \infty} (p_j^e, m_j^e) \neq (p^*, m^*)$. However, by Proposition 2, (p^*, m^*) is the

unique solution to $(p^*, m^*) = \arg \min_{p, m} \bar{AC}(p, \Lambda(m, 1))$ and the unique pair (p, m) associated to the allocation that maximizes steady-state surplus. That is, for possibly a subsequence, $\lim_{j \rightarrow \infty} S(p_j^e, n_j^e, m_j^e) \neq \mathcal{S}^*$, which is a contradiction. Therefore, it must hold that $\lim_{j \rightarrow \infty} \tau_j = 1$. \square